VALIDITY OF THE OVERCLAIMING TECHNIQUE AS A METHOD TO ACCOUNT FOR RESPONSE BIAS IN SELF-ASSESSMENT QUESTIONS.

ANALYSIS ON THE BASIS OF THE PISA 2012 DATA.

Thesis presented in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Sociology in the Institute of Sociology of the Jagiellonian University at Krakow, Poland,

BY

Marek Muszyński

Supervisor: Prof. dr hab. Jarosław Górniak

Kraków 2020

List of contents

List of abbreviations

2-PLM- Two-parameter Logistic Model (IRT)

AIC- Akaike Information Criterion

ARS- Acquiescence Response Style

BIC- Bayesian Information Criterion

BIDR- (the) Balanced Inventory of Desirable Responding

BPL- Bogus Pipeline

BTS- Bayesian Truth Serum

CFA- confirmatory factor analysis

C/IER- careless/inattentive responding

CIV- construct-irrelevant variance

DIF- Differential Item Functioning

ESDS- (the) Edwards' Social Desirability Scale

EEG- Electroencephalography

EFA- exploratory factor analysis

ERP- Event-Related Potential(s)

ERS- Extreme Response Style

ESS- European Social Survey

fMRI- functional Magnetic Resonance Imaging

GPA- Grade Point Average

GRM- Graded Response Model

IA- index of accuracy

IAT- Implicit Association Test

ICC- Item Characteristic Curve

ICT- Item Count Technique

IDR- ideographically desirable responding

IDR- Item Desirability Rating

IE- index of exaggeration

ILSA- international large-scale assessment

IM- (the) Impression Management

IRT- Item Response Theory

KAS- (the) Kwestionariusz Aprobaty Społecznej

LCA- latent class analysis

LSA- large-scale assessment

MCSDS- (the) Marlowe-Crowne Social Desirability Scale

MIRT- multidimensional Item Response Theory

MMPI- (the) Minnesota Multiphasic Personality Inventory

MRS- Middle Response Style, Midpoint Response Style

MTMM- multi-trait multi-method (model, matrix)

NLSA- national large-scale assessment

NPI- (the) Narcissistic Personality Inventory

OCQ- overclaiming questionnaire

ODQ- (the) Other-Deception Questionnaire

OCT- overclaiming technique

OECD- Organisation for Economic Co-operation and Development

PCM- Partial Credit Model

PFA- false alarm rate

PH- hit rate

PIAAC-Programme for the International Assessment of Adult Competencies

PISA- Programme of International Student Assessment

PV(s)- Plausible Value(s)

RRT- Randomised Response Test

RS- response style(s)

RSES- (the) Rosenberg Self-Esteem Scale

RT- Response Time(s)

S-C- self-consciousness

SDD- (the) Self-Deceptive Denial

SDE- (the) Self-Deceptive Enhancement

SDQ- (the) Self-Deception Questionnaire

SDR- Socially Desirable Responding

SDT- Signal Detection Theory

S-E- self-enhancement

TMS- Transcranial Magnetic Stimulation

TPB- (the) Theory of Planned Behaviour

UCT- Unmatched Count Technique

UIRT- unidimensional Item Response Theory

List of Tables

List of Figures

Abstract

The presented work is devoted to study the validity of overclaiming technique (OCT) as a measure of response (positivity) bias. Three main aims of the analyses performed were: a) assess methods' utility to enhance predictive validity of self-report by accounting for response biases, b) investigate proposed mechanisms of overclaiming, c) expand nomological network of the method by presenting a wide set of both individual-level and cluster-level (school) correlates.

The obtained results pointed that OCT can be used in order to account for response biases in self-report data. Important differences regarding use and interpretation of the different OCT scoring systems were found and commented. Two systems, one based of signal detection theory (SDT), other on item response theory model (IRT), were proposed as viable scorings of OCT. Choice between them is not trivial as it influences results' interpretation and model specification.

Three possible mechanisms of overclaiming were tested: a) motivated response bias (self-favouring bias, socially desirable responding), b) memory bias (overgeneralised knowledge or faulty memory control) and c) response styles and careless responding. The results pointed that all three mechanisms are probable and that overclaiming is most probably a heterogenous phenomenon of multiple causes. However, the analyses pointed out that one of the memory bias hypotheses, the overgeneralised knowledge account, does not hold and that there is much more evidence for the competitive metacognitive account. It is to said that overclaiming is at least partially attributable to insufficient monitoring of one's knowledge. Evidence for a relation between careless responding and overclaiming was also obtained, indicating that at least some of the overclaimed responses can be attributed due to inattentive responding. Obtained results on the relations between response styles and overclaiming were complicated; they warrant further studies as the results here probably greatly depend on the technical details of analysis, e.g. response style definition and coding adopted.

The analysed cluster-level covariates demonstrated that only very limited portion of OCT variance can be ascribed to the school-level of analysis. Gender, socio-economic status and locus of control proved to be significantly related to overclaiming among the individual-level correlates assessed. Boys yielded higher overclaiming bias than girls and students of external locus of control were more biased in their self-reports in comparison to students of internal locus of control.

The work comprises also analysis of the PISA's OCT latent structure. The results evidenced bifactor structure of the scale, with the general factor interpreted as math ability while the two specific factors were given a tentative explanation concentrated around item difficulty (one specific factor emerged for easy items, one for hard items). These findings point to a multi-dimensional character of OCT and a large role played by domain ability in OCT responding.

Moreover, latent class analysis (LCA) performed identified an "overclaiming" group among the participants which was characterised by high overclaiming and unwarrantedly high self-report profile regarding math-related abilities and social life. However, this group counted only around 9% of the total sample.

Implications of these findings are commented in the work, along with theoretical integration and ideas for future studies with the use of OCT.

Abstract in Polish (extended) [Abstrakt po polsku]

Przedstawiona praca analizuje trafność metody szacowania poziomu zawyżania oceny własnej wiedzy (*overclaiming technique;* OCT) jako narzędzia do kontroli i korekty efektu zbyt pozytywnego obrazu samego siebie (*positivity bias; self-favouring bias*) w badaniach samoopisowych niskiej stawki.

OCT jest metodą, gdzie sprawdzana jest deklarowana wiedza badanych w danej tematyce, np. w zakresie pojęć matematycznych. Do tego celu wykorzystuje się dwa rodzaje pozycji testowych: istniejące rzeczywiście pojęcia (np. „potęga") oraz pojęcia fałszywe, nieistniejące w rzeczywistości (np. „ułamek oznajmujący"). Przypisywanie sobie przez badanych wiedzy na temat pojęć fałszywych jest uważane za przejaw tworzenia zbyt pozytywnego obrazu samego siebie, co stanowi systematyczne obciążenie pomiaru i prowadzi do zmniejszenia trafności samoopisu. W pracy rozważany jest pomiar samoopisowy, gdzie respondent udziela informacji sam o sobie w reakcji na pytania kwestionariusza bez pośrednictwa ankietera (*self-administered self-report*).

Praca miała trzy cele:

- walidacja OCT jako metody kontrolowania efektów respondenckich w badaniach samoopisowych niskiej stawki,
- ustalenie mechanizmów prowadzących do przypisywania sobie nadmiernej wiedzy,
- zbadanie korelatów OCT.

Celem głównym było zbadanie praktycznej użyteczności metody przypisywania nadmiernej wiedzy do kontroli systematycznego błędu pomiaru w badaniach samoopisowych. Ważnym celem było również prześledzenie możliwych mechanizmów, które prowadzą do zawyżenia sądów na temat własnej wiedzy. Celem pomocniczym było zbadanie korelatów tego zjawiska, jako dodatkowego źródła informacji na temat możliwych mechanizmów zjawiska.

Najważniejszym elementem walidacji było ustalenie użyteczności wyników OCT do podniesienia trafności predykcyjnej miar samoopisowych oraz zwiększenia proporcji wyjaśnionej wariancji w modelu regresji, gdzie zmienną zależną był obiektywny pomiar umiejętności matematycznych (test PISA), a zmienną niezależną- wskaźniki zawyżania wiedzy.

Wyniki badania użyteczności OCT dowodzą, że metoda może być stosowana do kontroli błędów pomiaru w badaniach samoopisowych. Do oceny tego efektu przeprowadzono systematyczne badanie modeli supresji, które pozwoliły ustalić, że użycie miar OCT jako zmiennej kontrolnej prowadzi do zwiększenia trafności predykcyjnej samoopisu.

Przeprowadzone analizy doprowadziły również do ważnych wniosków na temat stosowanych do tej pory sposobów kwantyfikacji OCT. Wyniki prowadzą do wniosku, że używane niekiedy do tej pory miary nie powinny być stosowane i należy je zastąpić innymi wskaźnikami. Uzyskane wyniki pozwoliły na ustalenie, że miary oparte na teorii detekcji sygnałów (*signal detection theory*; SDT), jak również modelu odpowiadania na pozycję testową (*item response theory*; IRT) mogą zostać polecone jako wskaźniki OCT. Jednakże, jak wskazują przeprowadzone analizy, użycie jednej lub drugiej grupy wskaźników może pociągnąć za sobą uzyskanie odmiennych wyników, gdyż interpretacje obu grup wskaźników różnią się. Praca przynosi porównanie metod kwantyfikacji OCT i zwraca uwagę na kluczowe kwestie związane ze stosowaniem tych miar w modelach regresji i innych analizach ilościowych. Po raz pierwszy również zaprezentowano model interakcyjny miar, gdzie analizuje się moderację wskaźnika precyzji sądów przez wskaźnik nadmiernych sądów w ramach miar opartych na teorii detekcji sygnałów.

W pracy omówiono również trzy możliwe, sugerowane w literaturze, mechanizmy powstawania nadmiernych sądów o swojej wiedzy: a) motywowane zniekształcenie poznawcze, np. na skutek potrzeby aprobaty społecznej (*social desirability respodning*; SDR), b) niemotywowane zniekształcenia pamięciowe oraz c) przypisywanie sobie nadmiernej wiedzy jako produkt uboczny innych zniekształceń pomiaru (*response biases*), np. stylów odpowiedzi lub odpowiadania nieuważnego.

Zgromadzone wyniki wskazują, że motywowane zniekształcenia poznawcze mogą prowadzić do nadmiernych sądów, jednak wydaje się, że ten efekt nie jest silny i że takie efekty nie mogą wyjaśnić całej wariancji OCT.

Zniekształcenia pamięciowe wydają się istotnym kandydatem do wyjaśnienia wariancji zawyżania sądów na temat własnej wiedzy. Poczynione w pracy ustalenia wydają się wyraźnie sugerować, że efekty pamięciowe mają tutaj charakter błędów w monitorowaniu pamięci, a nie w nadmiernej generalizacji posiadanej wiedzy w skutek np. wzbudzenia rozbudowanej sieci semantycznej. Rezultaty wskazują, że zawyżanie sądów na temat własnej wiedzy lub umiejętności charakteryzuje raczej osoby o mniejszym poziomie kompetencji matematycznych zmierzonych testem kognitywnym.

Zawyżone sądy są również związane z odpowiadaniem nieuważnym, jednak dokładny wzorzec tego związku powinien zostać ustalony w kolejnych badaniach.

Analizy przeprowadzone w pracy zdają się sugerować, że metoda nadmiernych sądów ma heterogeniczną wariancję, na którą składa się kilka, niezależnych od siebie mechanizmów. Jakkolwiek zebrane dane nie pozwalają wnioskować wprost o zależnościach przyczynowo-skutkowych, to stanowią cenne źródło refleksji nad możliwymi mechanizmami zawyżania wiedzy przez badanych. Zgromadzona w pracy wiedza poszerzyła również informacje na temat relacji między odpowiadaniem nieuważnym, stylami odpowiedzi i metodami nadmiernych sądów. Praca jest pierwszym studium, gdzie użyto tak szerokiej palety wskaźników odpowiadania nieuważnego do analizy ich związków korelacyjnych z miarami OCT. Jest również pierwszym badaniem, gdzie użyto nowoczesnych, poprawnych metodologicznie miar stylów odpowiedzi do zbadania ich relacji z OCT.

Innowacją pracy jest również prześledzenie korelacji między normami społecznymi i innymi zmiennymi z poziomu szkoły a miarami OCT. Wyniki wskazują na marginalne znaczenie zmiennych szkolnych dla wariancji OCT.

Praca jest również pierwszym opracowaniem, które tak szczegółowo zbadało strukturę latentną skali zawyżonych sądów, doprowadzając do dopasowania modelu dobrze oddającego charakter danych empirycznych. Model ten wnosi interesujące implikacje teoretyczne, gdyż sugeruje, że badani nie odróżniają zbyt dobrze dwóch rodzajów pozycji testowych (prawdziwych i fałszywych), z których składa się metoda nadmiernych sądów. Rezultaty tego badania wskazują na kluczowe znaczenie umiejętności matematycznych dla przypisywania sobie zawyżonej wiedzy. Pokazuje ono również, że obiektywna fałszywość pewnych pojęć nie prowadzi do powstania jakościowo różnych mechanizmów odpowiadania na takie pozycje przez badanych. Wydaje się, że respondenci traktują takie pozycje jak trudne i mało znane pojęcia prawdziwe. Uzyskane informacje z analizy struktury czynnikowej OCT prowadzą również do praktycznych ustaleń na temat konstrukcji takich zadań w przyszłości.

Zgromadzone w pracy informacje i ustalenia istotnie powiększają wiedzę na temat samej metody, jak również możliwych mechanizmów zniekształceń pomiaru w badaniach samoopisowych. Praca integruje i porządkuje dotychczasową wiedzę o teorii i praktyce OCT.

Acknowledgements

First and foremost sincere thanks appertain to my supervisor **Professor Jarosław Górniak** who showed me the potential and beauty of quantitative skills and psychometrics which became my new and very enjoyed profession. Professor also gave me the idea to write about survey methodology for which I am grateful as it was a good choice indeed! I would also like to thank for supporting my grant proposals and schools abroad. I especially appreciate the motivational talks that encouraged me to work hard and set myself demanding goals. I am sure that if Professor had not become a scientist he would have made an excellent personal coach.

I would also like to thank all the people who provided me with their invaluable support in preparing and writing this dissertation. In alphabetical order I thank to: Francesco Avvisati, Krystian Barzykowski, Gabriela Czarnek, Paweł Grygiel, Grzegorz Humenny, Bartosz Kondratek, Hansjörg Plieninger, Artur Pokropek.

I am also indebted to a long list of wonderful, inspiring people that I have met during my university formation: supervisor of my MA thesis Professor Edward Nęcka, students and lecturers from the Institute of Psychology UJ, members of the Koło Naukowe Studentów Psychologii UJ (Marta Łukowska!), members of the amazing LangUsta (Psychology of Language and Bilingualism Lab) with Zofia Wodniecka, Jakub Szewczyk, Asia Durlik, Ania Marzecova and Michał Witkowski on top of the list here, students and staff of the Międzywydziałowe Indywidualne Studia Humanistyczne (MISH UJ), institution that gave me a five-year academic adventure and several lifelong friendships.

Special thanks appertain to all the participants of the Saturday PhD seminars- an exciting and fun place to discuss academic topics.

I also owe much to all of the students and lecturers of the Educational Measurement Studies, especially Paulina Skórska and Maciej Koniewski who put this extra effort to make the studies possible, and Steve Sireci, Chris Han and Maciej Jakubowski who were exceptionally inspiring and approachable teachers.

Much is owed to the staff of the reading rooms of the Jagiellonian Library, Książnica Podlaska and Biblioteka Główna of the Bialystok Technical University where I have wrote much of the dissertation and who provided me with all the help and materials I needed.

I would also like to thank all my Friends and family members who kept the fingers crossed for my success with a special *thank you* to my Grandmother who constantly asked me the “when” question. Well, Babciu, I think it is around “now” (finally!).

Most of all I would like to thank my Parents that have given me so much more that it is possible to fit in a piece of paper (even using MS Word), so I will simply write:

Mamo, Tato, dziękuję!

Chapter 1- INTRODUCTION

1.1 Key concepts and definitions.

The use of survey and questionnaire techniques is nowadays omnipresent in almost every area of social sciences (Paulhus & Vazire, 2007; Scandura & Williams, 2000; Ziegler, 2015), going also beyond the boundaries of this domain to medicine, technological sciences or biology (e.g. Fotios & Gibbons, 2018; Giromini, Viglione, Pignolo & Zennaro, 2019). Research conducted by Woszczynski and Whitman (2004) showed that around a third of articles published in top organizational psychology journals used self-report as their sole research methods, whereas Brutus, Gill and Duniewicz (2010) stated that self-report methods were used in around 60% of published organizational psychology studies. This popularity of the self-report methods is due to their cost-efficiency, ease of administration and flexibility to assess a broad range of constructs (Simsek & Veiga, 2001). Moreover, they are believed to provide valid, interpretative, standardized and comparable information across subjects (Lucas & Baird, 2004). This standardized information can then be easily translated into numerical data for further use in statistical modelling and formal testing of research hypotheses.

Because of that popularity it is essential that the measurement provided be as accurate as possible. Unfortunately, this method is subject to many errors threatening quality of the measurement and thus also of the inference. In a popular survey error classification Robert Groves mentions various potential sources of deviations (Groves et al., 2009). One of the prominent sources of error is the respondent himself/herself. Assumptions that respondents interpret, process and use the given response categories (options) in the same way (comparability assumption) and give unbiased and honest responses are not always held (Paulhus & Vazire, 2007; Wetzel, Böhnke & Brown, 2016). If this is the case it creates a deviation between the "true" level of a trait to be measured and the level actually measured (Groves et al., 2009). This deviance (gap) is commonly called "measurement error". However, this error can stem from many systematic and non-systematic processes. The non-systematic processes, called response variance, are based on a group of "haphazard processes", unpredictable over measurement occasions, bringing instability to traits' estimates (Groves et al., 2009). Nonetheless, there are the processes leading to systematic measurement error that truly pose a concern for every researcher employing survey questions. These systematic processes are jointly called "response biases", often defined as "systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content" (Paulhus, 1991) or "systematic distortion of a response process" (Groves et al., 2009). The main consequence of response biases presence in a survey or questionnaire measurement is a systematic under- or overreporting of a given trait level, thus, response biases introduce a systematic source of error variance to the measurement, reducing its validity and comparability (Podsakoff, MacKenzie, Lee & Podsakoff, 2003; Wetzel et al., 2016; Ziegler, 2015).

Numerous processes related to response biases have been identified, many of them stemming directly from respondents' traits or from an interaction of these traits and characteristics of a given research tool or measurement context (Ziegler, 2015). The response biases can be further divided into response styles (RS) and response sets (Cronbach, 1946; Jackson & Messick, 1958), the latter being a tendency that is to some extent dependent on particular measurement context and not generalisable to other occasions, whereas the former being defined as fixed, generalised disposition, that is stable over

measurement situations and even non-measurement behaviour (Jackson & Messick, 1958, 1962; Paulhus, 2002). It has to be noted however, that the differentiation on response sets and styles is not used universally by all researchers, despite the clear definitions provided by Cronbach (1946), Jackson and Messick (1958), Damarin and Messick (1965) and recently by Paulhus (2002). Sometimes only one of the terms is used, they are used interchangeably or response sets are seen as a type of response styles or *vice versa* (Groves et al., 2009; Weijters, Geuens, & Schillewaert, 2010)[1].

Sometimes a third type of bias is introduced, most often under the name of careless/insufficient effort responding (C/IER), also known under the terms of (pseudo)random responding, inattentive responding or inconsistent responding (Bowling & Huang, 2018; Curran, 2016; Fronczyk, 2014; Gibson & Bowling, 2019; Goldsmith & Clark, 2005; Johnson, 2005; Meade & Craig, 2012; Meyvis, Oppenheimer & Davidenko, 2009; Osborne & Blanchard, 2010). C/IER is at times included into the response styles framework (Loosveldt & Beullens, 2017; Van Vaerenbergh & Thomas, 2013) but in this work it will be treated as a third, distinct type of response bias. C/IER is defined as responding survey items without properly processing them cognitively, e.g. without reading the instructions or even the items themselves or reading the questions without full understanding (Kam, 2019). All of the response biases are believed to be inter-individually varied (Bolt & Johnson, 2009; John & Robins, 1994; Ziegler, 2015).

The processes laying on the basis of response biases are mainly researched under two frameworks. One of them is satisficing, defined as responding without putting maximum effort: respondents do only as little as they need to proceed through the survey (Krosnick & Alwin, 1988; Krosnick, 1991; 1999). The process is seen as a continuum between responding without any effort whatsoever (hard satisficing), responding with little effort (soft satisficing) and responding with optimal, required effort (optimising), sometimes even up to maximising (doing one's best). Although the framework is widely popular (the Krosnick's 1991 article has been cited almost 2000 times as yet) it is still believed to be somewhat under-researched, as the exact causes and correlates of satisficing (e.g. respondents' individual traits, research contexts, measurement tool characteristics and their interactions related to the degree of satisficing) are still elusive (Roßmann, Gummer, & Silber, 2017; Silber, Danner & Rammstedt, 2019).

The second framework is centred around the notion of socially desirable responding (SDR), which is defined as yielding overly positive self-descriptions, mostly by exaggerating one's positive traits or behaviours and diminishing negative ones (Paulhus, 2002). The given self-image is tailored according to the socially-appropriate characteristics (socio-cultural norms), as perceived by the respondent (DeJong, Pieters & Fox, 2010). The construct is often divided into various subtypes, depending on the theory there are two, three or even four forms of SDR postulated (Paulhus, 2002). Most often this concept is researched in two specific research situations: a) in high-stakes recruitment situations, where a particular type of SDR is often called "faking" (Griffith & Peterson, 2008; 2011; Ziegler, MacCann & Roberts, 2012) and b) in situations where respondents are to answer the so-called sensitive questions (e.g. questions about intimate or illegal issues) (Furnham, 1986; Krumpal, 2013; Tourangeau & Yan, 2007). These two situations clearly point to SDR as a motivated and deliberate process, but other researchers provide examples where SDR takes place without an apparent reason (at least external to respondent) and where it is considered a non-intentional process. Paulhus proposes a

[1] An example of such confusion is e.g. Rorer's (1965) proposal to name motivated distortions as sets and unmotivated as styles. This proposition was declined by many researchers on the ground that it advanced differentiations that could not be defended with the evidence-based knowledge on response biases (Messick, 1987).

framework of four SDR forms, two of them steered towards public impressions and two steered towards one's own view of the self (2002; see also Blasberg, Rogers & Paulhus, 2014).

Situations where research participants engage in SDR despite low-stakes and non-threatening measurement context are often considered a manifestation of a general cognitive bias of self-perception (overly positive bias of self-perception) (Bensch, 2018; Wojciszke, 2011; Ziegler, 2011). This overly positive view of the self (and often other elements of the world) is believed to have an adjustment component in maintaining well-being (Taylor & Brown, 1988, but see a meta-analysis by Dufner, Gebauer, Sedikides & Denissen, 2019 for a more nuanced view of this issue). However, there are concepts that has been differentiated from the overall positivity bias, as being less adaptive. One of them is self-enhancement defined as "exaggerating one's positive qualities" (Paulhus & Holden, 2009). There is an ongoing debate whether self-enhancement is a stable, trait-like tendency to evaluate oneself better than one really is, or is it only a state-like inclination, driven by research context or other transient factors (Robins & Beer, 2001; Taylor & Armor, 1996). Recent theoretical considerations incline towards stating that it is both, in a sense that there are forms of SDR that are context-dependent and there are forms driven by more stable motives (Paulhus, 2002). However, it is worth to note that self-enhancement operates also in contexts that are distinctive from the situations where SDR is most often researched (faking, high-stakes assessments and sensitive questions). It means that the research on SDR is often too narrowly focused, as this response bias is operational in a much larger number of research contexts, including the low-stakes one (which are typical for most of the situations in basic research context).

Thus, it can be summed up that there are three frameworks under which the response bias research is grouped: a) satisficing, where roots of bias are seen in respondents' low motivation or ability to participate in a survey, b) "classical" socially desirable responding, where participants are believed to deliberately manipulate their answers in order to adhere to socio-cultural norms or in order to make certain impression on others and c) "non-traditional" socially desirable responding, where participants' responses deviate from reality, without an apparent goal. The former SDR is believed deliberate, the latter, at least to some extent, non-deliberate and automatic. Sometimes differentiation on motivated and unmotivated cognitive biases is postulated, where "motivated", are understood as driven by intentional or unintentional motives, e.g. distorted self-perception, ego/self-esteem protection, etc. and "unmotivated" as stemming from other cognitive processes, e.g. memory biases (Muller, 2019; Muller & Moshagen, 2018, 2019a, 2019b). It is noteworthy though, that memory biases may also serve various self or social motives (Sedikides & Green, 2000), thus the boundaries of the unmotivated/motivated division of self-enhancement and other response biases are to be set more thoroughly with additional research.

Henceforth, any self-report measurement validity is threatened by many, often only partially ascertained, processes. Despite that many methods have been devised to control for response biases, to date no "golden standard" has been achieved so far. There are many methods proposed, some of them of dubious or unverified utility and almost all of them without a thorough validation. This situation is especially clear and enchaining in case of SDR in low-stakes, non-intrusive measurement situations where many methods developed for SDR prevention or control in high-stakes settings simply cannot be implemented. However, some new SDR control methods have been relatively recently proposed that are potentially available to capture this bias also in low-stakes measurement situations. Method that would to become truly useful for basic and applied research purposes, including international large-scale assessments (ILSAs), has to be easy to use and score and also pose low cognitive burden on participants. Moreover, it should be cost- and time-efficient and characterised by high level of flexibility, usable in diverse modes, populations and contexts. Most importantly, it has to

be a valid indicator of positivity bias, capable of controlling response bias variance and rising self-reports validity (Bing, Kluemper, Davison, Taylor & Novicevic, 2011; Ferrando, 2005; Ludeke & Makransky, 2016).

One of the most promising candidates for such a method is the overclaiming technique (OCT; Paulhus, Harms, Bruce & Lysy, 2003). The OCT is based on an idea proposed by Phillips and Clancy (1972), who asked their participants about their knowledge of some commercial products (e.g. books, movies, etc.), however, in order to control for social desirability tendencies, among the list of existent products, they have also placed some *non-existent* ones. In a list of famous politicians "Barack Obama" would be an example of an existent item, while "Peter Lawn" would be an illustration of a non-existent one. The terms "reals" and "foils" are often used to address these kind of items, respectively (Paulhus et al., 2003). Phillips and Clancy also proposed the terms "overclaiming" and "overclaimer". Overclaiming is thus defined as "overstating one's knowledge, by claiming to know non-existent items" (Muller, 2019). Obviously, apart from knowledge, participants can also assess their skills, abilities, behaviours, possessions, etc. The idea of this technique is simple- if participants claim to know non-existent items or to possess non-existent skills and abilities it is considered as a clear indication of their self-enhancing tendencies. Many solutions were proposed on how to score OCT data (and quite a few on how to use them in statistical analyses further on), but the most established method is based on signal detection theory (SDT), that offer a wide array of indices indicating participants' accuracy and penchant for biased responses alike (Paulhus & Petrusic, 2010).

The OCT has gained a sizeable popularity after Paulhus and his colleagues published their seminal article; since then, it has been cited almost 500 times so far, as indicated by the Google Scholar database. The OCT occurs in two main forms: a) questionnaires resembling general knowledge tests, like the one developed by Paulhus and colleagues (2003), called further the overclaiming questionnaire (OCQ), being close to the original idea of Phillips and Clancy (1972), where participants are asked about their knowledge or familiarity with a given set of items, b) tools resembling more a vocabulary test, like the vocabulary and overclaiming test (VOC-T) developed by Ziegler, Kemper and Rammstedt (2013), being a variation of lexical decision task paradigm (Meyer & Schaneveldt, 1971) or a similar task, vocabulary overclaiming in English (VOCE) developed by Dubois (2015). Among the first type of tools further differentiation for domain-specific and domain-general questionnaires can be proposed. An example of the former is the task developed especially for the PISA 2012 study (Kyllonen & Bertling, 2013), comprised of existent and non-existent mathematical terms embedded in the math concepts familiarity scale.

However, despite pegging the OCT as a very promising method to control for SDR in self-reports the results of its use are quite inconsistent. Along with corroborations of its utility to contain SDR variance (Bing, Kluemper, Davison, Taylor & Novicevic, 2011; Kyllonen & Bertling, 2013; Paulhus, 2011), many results question its usefulness in this context (Feeney & Goffin, 2015; Ludeke & Makransky, 2016; Kam, Risavy & Perunovic, 2015; Muller, 2019). The precise evaluation of the OCT's validity and utility to control for the unwanted error variance in self-reports is difficult, as among the numerous validity studies the majority of them was conducted in the faking paradigm, where measurement simulates high-stakes contexts, e.g. job application or college admission. The findings point that the context is not without the consequence on scores and validity of the OCT as some studies showed that OCT is a valid suppressor of SDR tendencies only in high-stakes, but not in low-stakes contexts (Bing et al., 2011). This and similar results question the technique's ability to contain for self-enhancement tendencies. Furthermore, even the technique's usability in controlling for impression management variance was challenged by the results provided by Feeney and Goffin (2015) and Ludeke and Makransky (2016), who found no relation between OCT bias index and typical measures of positivity

bias. The research to date shows that in certain contexts OCT can be a good method to control for faking, that is to say deliberate impression management leading to overly positive self-reporting of possessions, skills or knowledge (Dunlop et al., 2019). Nonetheless, the question whether OCT is capable of being a valid method to control for self-enhancement tendencies also in low-stakes contexts and in other paradigms than "faking" is still open to discussion. The current body of evidence does not offer any conclusive settling of this issue, thus warranting further research.

In order to plan a research aimed at answering this question there is a necessity to evaluate the theories regarding what processes are potentially engaged in responding to reals and foils. The accumulated research brought contending hypotheses on the mechanisms underlying embracing items in OCT. Traditionally, the technique was created in order to control for SDR, even before its various forms where recognised and named (Phillips & Clancy, 1972). Later on, it was seen mainly as a tool to account for deliberate reporting of an overly positive self-image, mainly in job-applicant or related contexts (e.g. Anderson, Warner & Spencer, 1984; Pannone, 1984), aiming at controlling the detrimental influence of what was named impression management (Paulhus, 1986; 2002). This kind of SDR is also often called faking, especially in the applied research context (Ziegler, 2011; Ziegler, MacCann & Roberts, 2012). However, controlling other form of SDR, self-deception, was also thought possible with the use of OCT (Randall & Fernandes, 1991) and this hope was explicitly stated and tested by Paulhus and colleagues (2003).

Nonetheless, in spite of the evidence heralding OCT's validity to account for both impression management and self-deception (e.g. Paulhus, 2011; Paulhus & Harms, 2004; Paulhus et al., 2003; Paulhus & Williams, 2002), there is also an ample evidence indicating that the high hopes related to OCT may not be fulfilled. The construct validity of OCT was routinely checked by correlating the technique with measures of various SDR forms. To this end questionnaires made to gauge SDR, the so-called SDR scales, were used, along with many techniques used to assess discrepancy between self-report and some more objective criterion (e.g. cognitive test, peer-report). Apart from measures of SDR and discrepancy also self-report scales of psychological traits related to SDR, e.g. narcissism and self-esteem, were commonly used. However, the result of these validity studies is mixed at best (Bensch, Paulhus, Stankov & Ziegler, 2017; Dunlop, Bourdage, de Vries, Hilbig, Zettler & Ludeke, 2017; Muller, 2019). Studies conducted by Feeney & Goffin (2015), as well as Ludeke & Makransky (2016) and Muller & Moshagen (2018) or Musch, Ostapczuk & Kleiber (2012) provided evidence questioning the believed validity of OCT to measure impression management and self-enhancement as OCT failed to correlate with measures of faking and discrepancy, as well as with self-report SDR scales, narcissism and self-esteem. These results put in doubt the whole nomological network of OCT and challenged its use as an SDR measure. Alternative explanations have been proposed, linking overclaiming origin to memory biases (Dunlop et al., 2017; Muller & Moshagen, 2019a) or careless responding (Ludeke & Makransky, 2016). Moreover, the overclaiming inter-individual variability has been never systematically evaluated, despite the evidence that SDR tendencies can vary greatly among the people (Leite & Cooper, 2010).

The nature of overclaiming is still little known and many pieces of important knowledge are missing. The technique has not been thoroughly tested so far, hence his work has been envisioned to provide additional evidence regarding the mechanisms of overclaiming, its nomological network and its inter-individual variability. As pointed by the newest articles in the field, these research problems still have not been solved adequately (Bensch, Maass, Greiff, Horstmann & Ziegler, 2019; Bensch et al., 2017; Ludeke & Makransky, 2016; Paulhus, 2011). This research is thus aimed to answer some fundamental questions regarding OCT as a measure of SDR and also to further the investigation of its mechanisms and correlates.

1.2 Research aim and objectives

The main aim of this research is to comprehensively validate the overclaiming technique as a method developed to control for SDR in self-report measures. This will be done by testing whether using OCT scores can account for spurious variance in regression equation.

The second aim is to extend the understanding of the method and refine its interpretation by comparing contending theories explaining its mechanisms of origin. Among the proposed mechanisms three will be tested here: OCT as result of memory bias (overgeneralisation of acquired knowledge or fault of metacognitive memory control), OCT as a result of positivity bias (SDR measure) and OCT as a result (by-product) of other response biases, namely response styles and careless responding.

The third aim is to present the overclaiming correlates in order to broaden knowledge about its nomological network. Apart from the individual-level correlates also the group-level (school-level) correlates will be explored in order to further understanding of OCT scores.

The detailed objectives of this research endeavour are to: a) validate the PISA's OCT and check whether it truly serves to control for SDR in low-stakes self-report research, b) expand the understanding of the method, by identifying the processes underlying its scores, c) search for its socio-psychological correlates, including group-level (school-level) correlates, d) investigate its inter-individual variability, e) compare and contrast various ways of scoring the OCT, f) assess latent structure of the analysed OCT version in order to inform theory on the basis of this findings and g) comment on the previous OCT research to make an integrative attempt at the state-of-the-art.

1.2.1 Research justification and significance

Self-report continues to be one of the fundamental research method in social sciences, due to its validity and research efficiency (Groves et al., 2009). However, its easiness to use does not come without a price to pay. This price is its proneness to faking and response biases. Already in 1960s Cook and Selltiz (1964) have pointed out that with self-report measures "the purpose of the instrument is obvious to the respondent; the implications of his answers are apparent to him; he can consciously control his responses. Thus a person who wishes to give a certain picture of himself-whether in order to impress the tester favourably, to preserve his own self-image, or for some other reason-can rather easily do so." In this statement Cook and Selltiz have pinpointed major motivated response biases, but it is important to remember that response biases can influence measurement also on the basis of nonmotivated, non-deliberate processes (Chambers & Windschitl, 2004; Muller, 2019; Muller & Moshagen, 2018; 2019a; 2019b). Top-notch quality measurement is the main requirement for social sciences advancement and its lack- "a major block to progress in sociological research" (Hauser, 1969).

Response biases pose a serious threat to any measurement's validity by introducing systematic error variance (spurious variance or construct-irrelevant variance; CIV; Messick, 1989) which often leads to distorted results and inference errors (Groves et al., 2009; Podsakoff et al., 2003; Schmidt, Le & Ilies, 2003). Leaving this variance uncontrolled can lead to many grievous consequences, all of them threatening both construct (Ziegler & Buehner, 2009) and criterion-related validity (Holden, 2007) and potentially influencing research conclusions.

One of the consequences is changed scales' dimensionality (Kam & Meyer, 2015). The factorial structure can be misrepresented, e.g. spurious factors (Huang, 2016; Woods, 2006) or higher-order, "method" factors (Bäckström, 2007) can emerge. These consequences often plague personality research, where positively- and negatively-formulated items (reversed items) regularly load on separate factors, despite the common underlying trait (Kulas, Klahr & Knights, 2018; Woods, 2006).

One of the most prominent examples of influence that uncorrected response biases had on research conclusions is the so-called general factor of personality- a higher-order factor emerging from personality scales, that has won plenty of research attention (DeYoung, Peterson & Higgins, 2002; Digman, 1997), but it is now believed to be rather an effect of measurement issues (Ashton, Lee, Goldberg & de Vries, 2009; Davies, Connelly, Ones & Birkland, 2015; Paulhus & John, 1998; Revelle & Wilt, 2013). Presence of error variance can also lead to infringement of the orthogonality (e.g. of factors or subscales) assumptions (Biderman et al., 2011; Khorramdel & von Davier, 2014). Measurement error is also, at least to some extent, responsible for the rising popularity of the bifactor scale structure, probably due to its potential to control for error variance (Reise, Kim, Mansolf & Widaman, 2016). Response biases can thus also distort model fit measures (Billiet & McClendon, 2000), influencing in this manner model choice decisions (Reise et al., 2016).

Another result are distorted means and variances, which can produce skewed distributions or spurious group differences, so that group means become uninterpretable (Chun, Campbell, & Yoo, 1974; Morren, Gelissen, & Vermunt, 2012). This potentially poses a critical problem in cross-cultural research and international large-scale assessments (ILSAs) (Bolt & Newton, 2011; Khorramdel et al., 2017). One of the common fallouts of SDR are elevated means (theta shift) of scales, often leading to paradoxical results, e.g. negative correlations between two, theoretically positively correlated measures (Kyllonen & Bertling, 2013).

The error variance can also lead to violations of measurement invariance and/or differential item functioning between groups of respondents (Bolt & Johnson, 2009), which is of paramount importance for any inter-group, especially cross-national, comparisons (Davidov, Muthen & Schmidt, 2018).

Another effect of error variance presence are deformed multivariate correlations, inducing spurious relations between the variables (Khorramdel & van Davier, 2014) or obscuring the true interdependencies between the measured constructs (He & van de Vijver, 2013; 2016; Lu & Bolt, 2015; Pokropek, 2014; Yang, Barnard-Brak & Lan, 2019).

Response biases can also deceive the commonly used measures of internal consistency (Fleischer, Mead & Huang, 2015), resulting in deflated or inflated measures, depending on the exact bias present in a given dataset.

Obviously, the response bias can influence research decisions and thus inferential errors of both I and II type (Johnson, 2005; Maniaci & Rogge, 2014).

Despite the long-realised presence of response biases and its dire consequences still no “golden standard” method to control for the response biases variance has been introduced in the field. Many methods have been proposed, but most of them can only be used in a narrow research context or is of still unverified validity (Wetzel et al., 2016; Ziegler, 2015). The proper discussion of response biases control and prevention methods and their vices and virtues is provided in one of the next chapters. Here it is however worth to note that among all the methods conceived to tackle with response biases, it is the overclaiming technique (OCT) that raised truly high hopes to be “the” method (Wetzel et al., 2016).

However, OCT still needs more research to verify its construct validity and usefulness as a method to control for self-enhancement tendencies. Moreover, inconsistent results obtained in the previous research attempts require disentanglement. What is more, precise interpretation of the method’s scores is still elusive as processes underlaying certain response patterns in OCT are not known. Additionally, the research presented here will bring more knowledge about the somewhat under-

researched OCT version, namely the PISA 2012 version of this method (to date the Paulhus' version brought much more research interest).

Moreover, the herein analyses will investigate the response biases problem in low-demand, low-stakes situations in a framework of educational studies, which is a context that is rarely researched in the SDR field. Furthermore, the analysed measurement tool type (self-assessment of educational skills) is also seldom investigated in the response biases framework, as most of the papers concentrate on personality scales, attitude and behaviour measures. Because of that the results of this work will be more readily applicable in a policy-oriented research, including international, large-scale assessments (PISA, PIAAC, ICILS,PIRLS/TIMSS, ESS, etc.), that comprise an important part of social sciences basic research. Response biases in the context of self-assessment of skills are far less researched, in contrast to the more readily applied investigations in the contexts of industrial, work and organisational psychology (IWO), psychopathology (forensic and clinical contexts) and sensitive questions (Krumpal, 2013; Perinelli & Gremigni, 2016) . As the recent results show, the response bias validity threat in these less-researched contexts is also serious (see also Burski, Chłoń-Domińczak, Palczyńska, Rynko & Śpiewanowski, 2013; Kyllonen & Bertling, 2013; Khorramdel et al., 2017; Pokropek, 2014), which seemed to be overlooked by many researchers (Meyvis et al., 2009).

Besides, the research presented here is based on a representative, random sample of 15-year-old high school students, which is markedly distinct from the most often employed non-random, convenient college students samples. Thus, the sample used in this work is larger from most of the other ones used in the field, is randomly-drawn and representative for a seldomly researched population (school students instead of college students or adults).

Further, the analyses will concentrate on the data from Poland, which is a rarely-researched culture in the context of response biases, as the field is dominated by the research on American samples (Dodou & de Winter, 2014). In this research only the data from Poland will be used in order to control for cultural and linguistical (questionnaire translation) effects (He & van de Vijver, 2016; Jerrim, Parker & Shure, 2019; Khorramdel et al., 2019).

The proposed research will also bring new evidence on the relation between overclaiming and response styles and careless responding, testing another hypothesis about the non-motivated origin of overclaiming. Moreover, this work proposes new interpretations of overclaiming, merging the methods and theories of various social sciences fields, mainly sociology and cognitive sciences. What is more, the analysis of contextual and group effects on overclaiming will be presented in order to explore possible relations between group characteristics (school-level) with individual-level overclaiming tendencies.

The proposed research is aimed to fill in the research lacuna by conducting a thorough validity study of the overclaiming technique and its utility to control for self-enhancement bias in self-report measures in the context of low-stakes educational research. The investigation is to examine the internal structure of the OCT questionnaire, its relation with the popular indices of careless responding and response style indices, explore its correlations with other self-report measures (e.g. math anxiety, math self-efficacy, etc.), and analyse contextual and group effects that may be related to overclaiming tendencies. All these aims are newly proposed, and to the best knowledge of the author, have not been explored to date or have been only very basically addressed.

Moreover, the research is also intended to conduct a criterion validity study of the OCT, looking to replicate its postulated role as a classical suppressor in a math familiarity-math ability relation (He & van de Vijver, 2016; Kyllonen & Bertling, 2013; Pokropek, 2014; Yang et al., 2019). This research is in line of the three proposed models of SDR effects in the data as: a) suppressor, b) moderator and c)

spurious correlation(s) (Ganster, Hennessey & Luthans, 1983). Furthermore, latent class models will be used to explore, whether the overclaiming tendencies are distributed universally across the whole sample or is it possible to distinguish subsamples that demonstrate large overclaiming or no such tendencies at all. Previous research on SDR and RS led to identification of such groups (Leite & Cooper, 2010; Khorramdel, von Davier & Pokropek, 2019), so it is warranted to explore the inter-individual variability of OCT and its correlates. Additionally, an analysis of relation of OCT to various RS and C/IER indices will be performed, in order to test the possibility that the overclaiming is a result of these processes (Dunlop et al., 2017; 2019). Finally, an analysis of different OCT scoring methods is proposed as an expansion of the analyses performed by Paulhus & Petrusic (2010).

To sum up: the proposed research is aimed to bring additional evidence on overclaiming technique utility in accounting for response biases in self-reports. Moreover, it aims to merge SDR, OCT, RS and C/IER research fields with the newest developments in social psychology, cognitive psychology and sociology. OCT is a very promising SDR control method, but additional information is needed to verify its utility, especially in front of evidence that threatens the view of OCT as "direct and unambiguous measure of an individual's attempt to deceive on a questionnaire (as the items are known to be non-existent)" (Randall & Fernandes, 1991).

Someone may question why PISA 2012 data has been used as a sole source of data in this work. However, this choice is justified and enables to fully answer the research questions posed. First of all, as it was already commented before, PISA offers a large, representative sample, larger than any study using any variation of OCT before. The sample size is not only a value in case of sample representativeness or statistical power, but it also enables use of more computationally-intensive methods of quantitative analysis, e.g. IRT models, which broadens the gamut of methods than can be used to appraise overclaiming features. The sheer size of the sample also helps to explore the inter-individual variability with the use of cluster analysis, mixture models or similar techniques- a given subdivisions of sample tend to be no larger than just a few % of the sample (Meade & Craig, 2012), which may be simply unobservable in small samples.

Moreover, the sample consists of high school students, which is a distinct sample from the two most often measured populations in the response biases research, namely college (psychology) students and online adult samples. Since the relation between overclaiming and age is not established (Clariana, Castelló & Cladellas, 2016) an opportunity to analyse overclaiming in this rarely researched population is most welcomed. Another related matter is that PISA comprises clustered data where students are nested within schools, which offers a unique chance to examine group and contextual correlates of overclaiming, a topic rarely and only preliminarily studied to date (Jerrim et al., 2019).

Furthermore, PISA offers data from many cultural contexts, including Poland, a rarely researched cultural context in the response biases field. As OCT is characterised by a significant cross-country variability (Vonkova et al., 2018) it is important to explore the characteristics of overclaiming in different countries, also outside the dominant Anglo-Saxon context (Jerrim et al., 2019).

What is more, PISA consists of many different self-report measures, an opportunity that provides chances for a very wide study of OCT correlates that is crucially needed to sketch technique's nomological network. Additionally, PISA grants an occasion to compare the validity of self-report by comparing it to an objective measure- PISA cognitive test of scholastic abilities. Having such a convenience is very rare in the discipline, but is highly valued as it enables a direct and valid test of OCT as a tool to measure positivity bias.

Obviously, this opulence of possibilities has been already explored by researchers in the OCT research (Fell et al., 2019; He & van de Vijver, 2016; Jerrim et al., 2019; Pokropek, 2014; Vonkova et al., 2018;

Yang et al., 2019) but these studies addressed somewhat different topics or in a different fashion than it is proposed in this work. What is more, some of them were only preliminary studies by far not depleting the analyses that can be performed with the use of such an ample database as offered by PISA 2012.

Finally, due to research effectivity it is better to make use of existing data sources to the maximum extend before plunging into collecting new data, especially, if such a large dataset is still underexplored in the context of OCT. It is also worthy to add that PISA 2012 is so far a unique occasion as OCT was not used in any other LSA since then.

1.2.2 Study design and research questions

Secondary data analysis was used in order to achieve the intended research aims. The plentiful PISA 2012 database was explored (Kyllonen & Bertling, 2013; OECD, 2014b). The instrument of interest is a math familiarity scale, consisted of 13 items, to which three additional items (foils) were added with an intent to measure overclaiming. The scale was implemented in the PISA student questionnaire as an indicator of opportunity to learn various mathematical concepts. This concept serves to directly assess students' exposure to different mathematical problems and to indirectly gauge instruction quality and students' abilities (OECD, 2014a). The OCT was used to "guard against the overclaiming" and to facilitate the inter-country comparisons in order to avoid the so-called "PISA paradox" (Kyllonen & Bertling, 2013). The "PISA paradox" is an effect where a given self-reported scale correlates with a different sign on individual, compared to group level of analysis. The effect is thought to be driven by response biases and social comparison processes (Khorramdel, von Davier, Bertling, Roberts & Kyllonen, 2017).

The math familiarity, as measured by self-report, should positively correlate with math ability, as measured by an objective cognitive assessment (PISA math test). Modelling overclaiming bias together with the level of self-described math familiarity should led to higher relation of the self-reported math knowledge with the actual knowledge, as measured by the test, thus showing that OCT scores can indeed account for response bias variance. To perform this analysis a multilevel linear regressions were used. These analyses constitute a criterion validity study of OCT as a method to control for self-enhancement bias.

Moreover, the internal structure of the math familiarity scale was thoroughly examined by confirmatory factor analysis (CFA) in order to test and verify theoretical assumptions underlying the math familiarity scale and the overclaiming technique. This analysis is also meant as a validity study, investigating the internal structure as a source of construct validity.

Latent class analysis (LCA) models were implemented to search for latent subsets of participants, characterised by different responding to the math familiarity (and OCT) scale. This method is an expansion of the above-mentioned construct and criterion validity studies; it also gives insight into the inter-individual variation of self-enhancing tendencies.

Moreover, an in-depth OCT convergent and divergent validity study was undertaken. Among others, OCT correlations with various careless responding (C/IER) and response styles (RS) indices were examined. The IRTree framework was used to construct and calculate the RS indices (Böckenholt, 2012; Khorramdel & von Davier, 2014). This analysis will also shed light on the mechanisms of origin of the overclaiming, being thus another construct validity test.

As a continuation of convergent/divergent validity studies the correlations between OCT and various socio-economical, psychological and educational measures included in the PISA 2012 study were

assessed. What is more, most of these concepts were analysed on both individual- and school-level to account for individual, contextual and group effects (Raudenbusch & Willms, 1995).

1.3 Dissertation plan

Chapters two and three of this dissertation presents the theory and history of the positivity bias research with a focus on self-enhancement and SDR. A succinct review of complementing theories and related concepts (self-consciousness, self-knowledge) is also presented.

Chapter four introduces research concentrated on the validity of self-report, followed by thorough overview of the methods used to control for SDR variance.

Chapter five reviews the state-of-the-art in the OCT research field with a special focus on validity studies and attempts to explore OCT mechanisms. Precise research hypotheses are verbalised in this chapter as they evince from the questions and doubts formulated in the literature. Research lacunae are identified and ways to fill them in are proposed.

Chapter six comprise a general method section, where methodological issues common for all research questions are described. A description of the database used and a general overview of PISA research is also included here.

Chapter seven presents the analyses aiming to solve the research questions posited in Chapter five. Each research question has its own subchapter, consisting of descriptions of methods and results with a discussion as conclusion.

Chapter eight offers general discussion, where all research questions results are commented, synthesised and their implications for the field are presented. Future directions of further research are proposed here, along with limitations of the herein analyses. All of the above chapters are followed by a succinct chapter summaries (save Chapters 6 and 7 that do not need them).

The work is concluded with a list of references and technical appendices.

1.3.1 What this work is NOT about?

At the end of this part it is also important to mention what this work will not be about. Issues like response and wording effects (Bradburn et al., 1979), sensitive questions (Tourangeau & Yan, 2007), mode comparisons (Dodou & de Winter, 2014), interviewer frauds (Kemper & Menold, 2014), malingering or faking bad (Rogers, Gillis, Bagby & Monteiro, 1991) and SDR in clinical (Logan, Claar & Scharff, 2008) or forensic settings (Tan & Grace, 2008) are completely out of the scope of this work.

Problems regarding cultural differences in response biases will be mentioned only when directly relevant to the main topic of the dissertation. A preliminary study of cultural differences in OCT is offered by Vonkova, Papajoanu & Stipek (2018) and Fell, Koenig, Jung, Sorg & Ziegler (2019), whereas a thorough discussion of RS and their cultural differences is presented e.g. in the works of He and van de Vijver (2015a; 2016), Johnson, Kulesa, Cho and Shavitt (2005), Ju and Falk (2019), Peterson, Rhi-Perez and Albaum (2014) and Van Vaerenbergh and Thomas (2013). One of the first studies (the first?) exploring the relations between culture and careless responding was recently published by Grau, Ebbeler and Banse (2019).

Furthermore, research problems related to overclaiming, but distinct from it like e.g. overconfidence (Koriat, Lichtenstein & Fischhoff, 1980), feeling-of-knowing (Hart, 1965; Thomas, Bulevich & Dubois, 2012), bullshitting (Frankfurt, 1986), pseudo-opinions (Bishop, Tuchfarber & Oldendick, 1986) and Kruger-Dunning effect (Kruger & Dunning, 1999) will be related to only when directly relevant to the

main topic of the dissertation, but will not be reviewed and commented in the whole extent. Other methods of SDR correction that are not used to validate OCT will be treated in a similar fashion. Their comprehensive reviews can be found elsewhere (Franzen & Mader, 2019; Furnham, 1986; Hipsz, 2014; Krumpal, 2013; Nederhof, 1985; Paulhus, 1991; 2017).

1.3.2 Terminological note

At the end of this part a small terminological note is in place. In the field of response biases many concepts of unclear or overlapping meaning are used without sufficient specificity. However, in this work, whenever referring to the contemporary state-of-the-art terms like “faking”, “impression management” and “self-presentation” will be used to denote process of creating overly positive self-image in order to achieve certain goal. As Ziegler defined it: “faking is an interaction between person and situation resulting in disrupted item responses aimed at a personal profit for the testee”. In this view faking is understood more as an evolutionary necessity, game theory consequence (“if I don’t fake, I’ll lose, because everybody else is faking”), camouflage, deception- an inherent part of life, that is short to blatant lying and manipulating (Depaulo, Kashy, Kirkendol, Dwyer & Epstein, 1996; Griffith & McDaniel, 2006; John & Hogan, 2006; Smith, 2004). This bias is considered a threat only in high-stakes contexts as in low-stakes assessments there are no profits to be gained. Other typical situations where faking is an issue are sensitive questions and similar types of contexts. In this occasions respondents may resort to faking in order to conceal their true answers or impress an interviewer.

However, in low-stakes conditions other kind of bias is more probable the outcomes of which are very similar to faking- an overly positive image of respondent. In this work it will be called positivity bias or overly positive bias as this is a broader and more precise term than previously used socially desirable responding or self-enhancement. However, when referring to historical conceptions and results the terminology used by their authors is mainly kept.

Regarding other terms, names like self-evaluations, self-ratings, self-assessments, self-descriptions are treated as close synonyms, whereas words pertaining to certain theories are used in the senses denoted by those conceptions. In this group are terms like: self-enhancement, self-evaluation, self-image, self-verification, self-confirmation, self-knowledge, etc.

THEORETICAL PART: SOCIO-PSYCHOLOGICAL BASIS OF POSITIVITY BIAS AND OTHER RESPONSE BIASES IN SELF-REPORTS

Chapter 2- SOCIALLY DESIRABLE RESPONDING

2.1 History of the concept

2.1.1 Philosophical and socio-psychological background

The concept of self-knowledge and false judgments about oneself were already known in the antiquity. Famous Greek statesman and orator, Demosthenes, stated: "Nothing is so easy as to deceive oneself; for what we wish, we readily believe", whereas the Delphian temple had the inscription "know thyself" engraved above the entrance. This maxima was also adopted by Socrates as a fundamental and most important task laying before everyone. Knowing oneself would, according to Socrates, lead to knowing also the truth about human being as well as understanding others. Independently from ancient Greeks this line of thinking was also popular in the Far East, where philosophical system constructed by Lao Tzu equalled knowing oneself to "enlightenment", the highest and most difficult knowledge to attain. Already from these early beginnings of scientific inquiry the philosophers argued not only about the nature of the self but also on the sureness of what one can know about oneself. Differentiating of what was real and what only apparent was one of the cornerstones already of the pre-Socratic philosophy. Plato, in his famous cave example, claimed that we will never get to the exact nature of the things as we can only observe the shadows of the ideas. Many years later, Descartes was also sceptical about the possibility to know the things, one thing he was certain of, was the independence of self-awareness from any physical experiences, senses or matters (this concept dates back at least to Avicenna) and by the sole fact of being self-aware one can state that he/she exists. Hume, however, was not so certain that we should speak about one "self" as he believed that each and every person is constructed by different "selves" constituted by heaps of different, everchanging thoughts, judgements and sensations ("perceptions") and that "self" is nothing more than a "bundle of perceptions". Daniel Dennett went even that far as to state that no self actually exists, it is just a "narrative centre of gravity", part of the world sensemaking process (1992). On the other hand, William James proposed an integrative theory of the self by stating that it can be divided into "I" and "Me". "Me" could be further divided into material, social and spiritual self. All three of them were changeable in the course of life. The material self is related to one's body, known people and possessed things. The social self is what determines behaviours in every social situation and according to James this changes from a situation to situation. The spiritual self encloses subjective knowledge about one's personality, values, etc. All three "selves" are integrated by the "I", an indivisible entity that integrates and unifies all past, present and future experiences into one coherent unity (1890).

At first philosophers were quite assured about the certainty of self-knowledge due to its proximity to self and ease of its "measurement" in the way of inner observation process known as introspection. Moreover, all self-related actions and attitudes were perceived as conscious, for being conscious of one's acts was one of the main characteristics of self-awareness, postulated e.g. by Descartes or Locke. However, empirical evidence brought an end to this optimistic accounts. As reviewed by Nisbett and Wilson (1977) introspection and self-inference are prone to fallacies and biases and are rarely capable of creating "generally correct or reliable reports". This evidence was however criticised as limited only to certain mental states and methods of reaching self-knowledge (Wilson & Dunn, 2004). Nowadays, there is only a general consensus about the methods in which the self-knowledge can be acquired. Boghossian (1989) noted three such methods: a) introspection, b) (self-)inference, and c) (self-)verification by simply thinking about one's states or thoughts. Other methods, more established in the empirical knowledge, included: a) introspection, b) looking glass self- inferring about one's traits on the basis of how others behave towards us, c) social comparisons, d) self-perception- inferring about oneself on the basis of one's behaviour (Baumeister & Bushman, 2011). However, the once-

postulated privileged access to one's thoughts seemed to be an overoptimistic account as soon as researchers became aware of various biases and limitations of self-knowledge (Taylor & Brown, 1988). Moreover, it was discovered shortly afterwards that people not only conceal their true selves from other people but are also good at self-deception. Altogether forming a reliable, coherent image of oneself is not an easy task. As it was wittily stated by Benjamin Franklin: "there are three Things extremely hard: Steel, a Diamond and to know one's self".

Despite of these difficulties and reservations "knowing oneself" became a key skill required from participants of early psychological and sociological research. The advent of the self-report method in social sciences is dated at least from 1917 and the first personality scale (then named "data sheet") introduced by Woodworth[2]. Soon, the self-reports entered also other research fields, including attitudes, behaviours, knowledge and skills measurement (Thurstone, 1928), becoming one of the dominant research methods whenever human subjects were considered. However, the researchers soon became preoccupied by potential response distortions (biases) and the threat to the validity of measurement they posed. One of the main troubles was a "tendency to give answers assumed to be socially approved" and "being warped by the subject's desire to be other than he really is" (Bernreuter, 1933). Other researchers were concerned by "faking", understood as a lack of "honesty and sincere cooperation" from the part of the respondent (Allport, 1928). Frenkel-Brunswick (1939) differentiated between sincere and insincere self-reports ascribing the latter to "self-deception" or "auto-illusions". Comparing the self-reports with judges opinions on participants' behaviour, she noticed that participants omitted certain characteristics, characterised vices as virtues, justified the defects and minimised their importance among many other distortive tendencies in self-report. Frenkel-Brunswick ascribed the self-deceptions to social maladjustment and personality defects, in some cases even to psychopathology, and observed that they were correlated with personality traits, but not with intellectual ability (1939). She believed that self-deception may perform a role of defensive mechanism, being "comforting and helpful", but also noted that responding to deficits with self-deception may bring both good and bad consequences. The former from successful "impressive mechanisms", helping adjusting to the environment, the latter from "keeping the individual unaware from his shortcomings", namely, successful self-deception that prevented reading any feedback from the society. Obviously these views were borrowing heavily from Freudian theory that self-insight was limited by the defensive mechanisms which served to keep negative and harming information away from burdening the conscious self (Freud 1938/1941). Unfortunately, Frenkel-Brunswick's research could not be continued due to the Anschluss that forced her to leave her native Austria and flee to the USA where she picked up different research topics, much influenced by her pre-war European experiences[3].

The above-mentioned studies were among the first where the threat to measurement validity from "self-deception" or "faking" was named and explored. However, the early attempts of Frenkel-Brunswick and others to investigate the correlates and mechanisms of this phenomenon were largely replaced by a pragmatic quest to find the best method to capture and control the bias. This atheoretical approach (Zickar & Gibby, 2006) caused much trouble to the discipline in the following years. However, back in 1930s, as the personality testing business was growing, the researchers were urged to find a

[2] I consciously omit here the very early attempts of Wundt and Titchener as they were not addressed by the core researchers in the response bias field.

[3] Her main collaborator- Shmuel (Siegfried) Nagler- also had to flee the post-Anschluss Austria. More on his story: https://gedenkbuch.univie.ac.at/index.php?id=435&no_cache=1&L=2&person_single_id=1768&person_name=&person_geburtstag_tag=not_selected&person_geburtstag_monat=not_selected&person_geburtstag_jahr=not_selected&person_fakultaet=not_selected&person_kategorie=not_selected&person_volltextsuche=&search_person.x=1&result_page=90

method to pre-empt response biases or at least control them in order to prevent them from distorting the scores of the best-selling measurement tools, widely applied in industrial, work and organisational contexts, e.g. in employees selection. This urge was especially fuelled by the research findings of e.g. Steinmetz (1932), Bernreuter (1933) or Vernon (1934) that indicated, too much woe of the researchers, that participants are both motivated and able to manipulate their responses in order to make certain impression or simply to achieve high score enabling to obtain a job. Apart from selecting or promoting best suited applicants the early response bias research was also obsessed with screening out the maladjusted, disturbed and potentially delinquent candidates/workers (Gibby & Zickar, 2008) that could potentially lied or try to "hack" the measurement in order to avoid identification. Another field of research that was much preoccupied with participants' dissimulation was clinical and forensic psychology that faced a serious threat of using distorted data in their practice. Hence, the early research endeavours identified two types of response distortions: a) deliberate (often referred to as "conscious") tendencies to yield untrue self-image, to deceive others on purpose, most often called "faking" or simply "lying" and b) an unconscious (or less conscious) penchant for presenting oneself in a better light, most often named "self-deception" or "defensiveness", which role was believed to protect self-esteem and self-beliefs (Paulhus, 1986). However, no established methods were available at hand to control for these biases in self-report data.

2.1.2 Early methods to control for SDR bias

Therefore, as there was no Delphian oracle nearby to aid them in knowing the participants better, the early researchers plunged into conceiving the best methods to control for response biases in self-report data. One of the first attempts was taken by Hartshorne and May (1930) who were devoted in a basic research of finding character (personality) traits that were related to lying and created a self-report scale that was devised to identify dishonest responders. The researchers also laid the foundations for the dual perception of dishonest responses: as stemming from trait-like characteristics of respondent but also originating from "transient needs and opportunities to deceive", created by the measurement situation (cf. Ziegler, 2015). This approach was to become a prototype in the field- to create a specially tailored self-report scale that, unknowingly to the respondent, would reveal any distorting tendencies in her responses. This kind of self-report would later on be known as "lie scales". This was precisely the name given to another such instrument, created by Hathaway and McKinley (1943) in order to control for faking in personality self-reports used in clinical settings. This scale was soon succeeded by the so-called "K scale", which role was to "detect (…) defensiveness" (Meehl & Hathaway, 1946). The scale was constructed by picking up items that best discriminated between subjects diagnosed as abnormal vs. normal by clinicians, but that had normal profiles from the personality tests. Moreover, the authors of the scale wanted that the scale itself did not capture any specific construct but instead measured only pure response bias. In their own words: "[it] was not assumed to be measuring anything which in itself is of psychiatric significance" (Paulhus, 1986). The K scale, along with the L scale (for Lie scale), soon became part of the then most popular personality questionnaire, the MMPI (Minnesota Multiphasic Personality Inventory), constituting one of the methods' validity scales, included to control for response biases (Hathaway & McKinley, 1951). Thus, the MMPI entered the research practice guarded by several validity scales: K, measuring self-deception tendencies, L, capturing faking good and F, which was an interesting innovation in the MMPI intended to measure faking bad (malingering) tendencies.

The approach consisting in creating special scales to identify distorted response profiles became very popular henceforth and it will be addressed again in this dissertation in the review of contemporary methods to control for response biases (mainly in the subchapter 4.5).

A somewhat distinct method was used by Humm and Humm (1944) in order to control for the "bias of frankness" in the validity studies of their personality scale HWTS (Humm-Wadsworth Temperament Scale). Their method was named the No-count as it was based on simply counting the number of "No" responses among the 318 items of the questionnaire and comparing the count with some quite arbitrary norms of what number of "No" answers was deemed acceptable or unacceptable by the authors of the scale[4]. In fact this approach was quite similar to the one adopted in the MMPI, the main difference was that Humm and Humm advocated to count the number of certain responses in the same measurement tool that was to be shielded from distortion, whereas the MMPI approach was based on adding additional tools (and quite long) to the proper construct scale. Humm and Humm later on proposed also additional uses of the No-count technique: plotting profile scores and counting a deviation of a given score from the mode of scores in a given scale (1947).

Different method to control for response biases was proposed by Ruch (1942). It was based on comparing two scores- one when participants were instructed to respond honestly and one were told to respond as if they were applying for a real job. By comparing the two scores Ruch was hoping to identify the fakers. However, the method soon revealed its problems- it was difficult if not impossible to distinguish someone truly high on a given trait, from a distorted profile achieving the same high score (Zickar & Gibby, 2006). Despite this inherent limitation the Ruch's technique gained significant popularity and will be also referred to later on in the work (subchapter 4.5).

2.1.3 Response sets concept

Apart from faking and self-deception also a third type of response distortion was identified early on in the literature- these were the so-called "work methods" (Seashore, 1939) and "response sets" (Cronbach, 1946). Seashore defined the former as individualised patterns of behaviour adopted during learning or during solving a problem such as a psychological test (including personality measures under "test"). He believed them to be a source of significant inter-individual variability and called to control them in psychological measures. Cronbach defined the latter as "any tendency causing a person consistently to give different responses to test items than he would when the same content is presented in a different form". Despite the different terms and definitions both researchers have thought about the same idea- that according to the measurement theory the content of an item should be the only source of variability. However, it is not the case, as many other sources exist, e.g. item form, measurement context, individual differences in responding techniques, all have its influence on score variance. Seashore argued that the very same respondent can achieve very different scores depending on the responding technique used (1939). Seashore stated that mainly thinking about performance tests (e.g. motor skills measurement), but Cronbach adopted his way of thinking and generalised it also to self-reports (1941)[5].

[4] Scores of 194 or more "No" answers or 144 or less of them were considered unacceptable (Humm & Humm, 1944). It seems that the numbers were just loose criterions set on the basis of scores distributions. On the basis of this it can be inferred that the expected number of "No" answers was estimated around 50% of the total number of answers.

[5] Probably the first work that pointed out that other factors than content of items were influencing responses in self-reports was Thorndike, who in 1920 described his works on the halo effect. He discovered that ratings on seemingly unrelated traits were correlated, indicating that participants used some kind of response technique, e.g. they inferred on a trait from other traits, or used an overall impression to rate every other detailed aspect of a person. Thorndike was thinking mostly about the physical appearance as the basis of all other ratings, but his discovery was further developed by Rundquist & Sletto (1936) and Lorge (1937) towards understanding halo effects in a response set manner. Applying acquiescent response set would also result in correlations observed by Thorndike (Cronbach, 1946).

Among the response sets distinguished by Cronbach were: speed vs. accuracy, caution vs. incaution/gambling, acquiescence, differences in defining response categories (especially vague terms like “desirable”, “frequently”, etc.) and inclusiveness (1946). The last two effects are now classified as response effects, variations in self-report responses resulting from minor characteristics of measurement tool form, design or administration, as it was stated above, these group of biases will not be described in detail in this work. On the other hand, the other response sets described by Cronbach are mostly comprised within the response bias framework and often researched together with faking or self-deception topics (Zickar & Gibby, 2006). Acquiescence, defined as tendency to endorse positive response options (e.g. “true”, “I agree”, “right”, etc.) independently of item content, gained a lot of research attention henceforth (Baer, Rinaldo & Berry, 2003)[6].

Cronbach’s another important contribution was the question about the response sets stability- are they only “temporary sets” or could they be more stable “habitual techniques of performance”? This differentiation between transient and stable character of certain response biases proved to be an important differentiation in further research attempts. Cronbach himself found evidence for both temporally stable and changing response sets, pointing the latter as the most serious validity threat of all the response sets types (1950).

Similarly important question was whether the response sets are only “incidental sources of error” or whether they reflect meaningful traits? Cronbach observed that to a minimal extend the response sets were correlated with “external variables such as attitudes, interests and personality” (Cronbach, 1950). He even called response sets a paradox, as they are a convolution of meaningful variance and measurement interference. However, this character of the response sets may be the result of the imprecise nature of the methods used to identify them, hence Cronbach urged the quest to devise “a pure measure of the response set itself” (1950). Along the methods proposed to prevent the response sets (e.g. tailoring test design, response format, instructions form, etc.) were also methods meant to correct for response sets, but separating the “constant error” from trait variance proved to be a very difficult task. Cronbach argued that only such pure measures could be used as a suppressor variables, which would be an ideal way of separating the variance sources in a regression equation.

This call was in general backed up by Jackson and Messick (1958), however, they have also proposed some alterations to the Cronbach’s concepts. They have suggested to change an ambiguous "set" to a simpler "style" and viewed "style" as a potential expression of construct-relevant characteristics, not only as a measurement nuisance. They have also criticised the existing measures as confounding style with content, hence not offering a clear measurement of neither. Jackson and Messick opposed views of Gage, Leavitt and Stone (1957) who claimed that response sets are fortunate, because oftentimes they lead to higher criterion-related validity, by pointing that response sets compromise construct validity: "conglomerate indices containing both content and style will not suffice and will confuse the issues". Jackson and Messick (1958) also advocated the development of new response sets/styles measures on the basis of theoretical advancements and criticised the then-popular data-driven approach. Finally, they have indicated that various response sets/styles can be related to each other and can even interact, as it was often postulated in case of acquiescence and SDR. They research also

[6] The effect is believed to be discovered by Martin F. Fritz in 1927, who noticed that in a true-false medical knowledge test with a balanced number of correct true and false answers the respondents chosen true response option in 62% of the items. Fritz concluded that the guessing “was not equivalent to the tossing of a penny”. Fritz, however, did not use the term “acquiescence” in his publication. It was probably brought to the social sciences (in this sense) only by T.F. Lentz in 1938.

paved way to the later conceptions seeing SDR more as a personality trait than a measurement artefact.

It is worth to note, that Jackson and Messick's article also brought some definitional confusion as many articles used the terms "response sets" and "response styles" interchangeably, without much reflection upon their intended meanings. The matter was clarified by Paulhus (2002), who provided clear-cut definitions terming "response styles" as "biases that are consistent across time and questionnaires" (according to the Jackson and Messick's view of response style as an *á la* trait entity) and "response sets" as "short-lived biases attributable to some temporary distraction or motivation" (in relation to Cronbach's idea of sets as measurement errors). Paulhus also added elsewhere (2003) that when a distortion is stable across time and measurement contexts and has its own "cognitive and/or motivational roots" then "they can be studied as personality traits in their own right (…) with their importance going well beyond the self-report measures" (see also Wetzel, Lüdtke, Zettler & Böhnke, 2016). Therefore, there are two dimensions around which response sets and styles can be defined and differentiated: temporal status and relation to substantial, construct-related variance. Response styles are defined as stable and related to some substantial, trait-like variability, whereas response sets are believed transient and trivial, conveying only error, construct-irrelevant variance.

2.1.4 Edwards' conceptions and their critique

As a way to find better methods of response sets corrections Cronbach saw item-level analysis, pointing out that item is not constant for every person but that the response is to the item and it is a function of inter-individual variables and item characteristics. Among the key item features driving response set answers were item difficulty and vagueness (1950). Edwards continued with this line of reasoning and added another item feature critical for the validity of the scale- item social desirability (1953). Edwards provided evidence that the item endorsement is strictly correlated with its perceived social desirability. His point of departure was an observation that traits' social desirability may be responsible for high correlations observed among personality scales (1957). In his further research attempts he managed to accumulate evidence that social desirability responding (SDR) could be qualified a response set *sensu* Cronbach, but also considered that this tendency can be "a fairly stable personality characteristic" (Edwards, 1957).

He also commented that what participants really perceive as socially desirable need to be thoroughly checked, as it tended to change across groups (e.g. professions) or across cultures, thus comparing a trait endorsement in a given number of groups, one should first ascertain that the measurement tools' items have similar social desirability in all of the groups. Otherwise, it would be impossible to determine whether the differences stem from the difference on trait level or the difference in items' social desirability (1957). Moreover, Gordon (1951) stated that within-group trait desirability is correlated with the application of this trait to a given participant, on the basis of which the popularity of this trait in a given group is determined and it is from this popularity that the social desirability of a trait is determined[7].

Edwards thought that socially desirable responding is present in both faking and self-deception and came up with a scale measuring SDR tendencies in self-reports. To this end he mixed the items picked from several MMPI scales (K, L, F and Manifest Anxiety(MAS)) that were assessed as most socially desirable by a set of judges. From these items he formed a new scale to measure SDR tendencies which

[7] In other words: a medical doctor perceiving herself as intelligent would claim that in her professional group there are a lot of intelligent individuals and that intelligence is highly valued (socially desirable) among the physicians. Someone with different self-perception might have different opinion on this matter.

he called Edwards' Social Desirability Scale (Edwards' SDS; 1957), providing yet another measurement tool to control for response distortions in self-reports.

However, the validity of this new measure was quickly questioned by Crowne and Marlowe (1960). These two researchers criticised the foundations of the Edwards' scale, namely the method used to pick the items from the whole MMPI. Selecting the items from an inventory intended for clinical purposes put in doubt the interpretability of scores in case of normal populations. Many items had clear psychopathological, abnormal implications, e.g. sleep problems. Hence, a researcher could never be sure, whether respondents denying such items in fact did not suffer from these problems or whether they were responding desirably. This meant that the scale's score would be always a confound of the socially desirable responding tendencies and psychopathological symptoms. Crowne and Marlowe have hence criticised the Edwards' SDS content validity as it defined SDR only in a very narrow sense of denying or accepting maladjustment and clinical problems. To solve these issues they have proposed creating a new scale without drawing the items from the MMPI. A new set of items was prepared with a main target to contain socially approved and culturally sanctioned behaviours which were, however, "improbable of occurrence". These items were meant to be without any psychopathological implications and to measure SDR independently from clinical traits. Crowne and Marlowe have confirmed certain independence of their scale by achieving lower correlations with various MMPI scales than were achieved by the Edwards' SDS (1960). By preparing this scale the researchers have provided a less extreme scale, better suited for general populations. Also, a new conception of SDR was put forth as it was defined as need of social approval, "the need of subjects to respond in culturally sanctioned ways", thus Crowne and Marlowe were playing down the explanations like faking and self-deception, paying more attention to the social context of measurement and group norms (1960). Their scale, named Marlowe-Crowne Social Desirability Scale (MC-SDS, written also M-C SDS or MCSDS), was to become the most popular SDR scale for the next 25 years (Paulhus, 1986).

However, the research on the dimensionality of the SDR scales and other "stylistic" measures brought significant doubts regarding the conclusions reached by Edwards (1957) and Crowne and Marlowe (1960). Messick (1960), using factor analysis, provided evidence that SDR scales and similar instruments did not converge to one dimension, but instead formed a lot of different item clusters. Some of them were small, consisted of item doublets or triplets organised around a semantic issue, but some of them seemed legitimate factors representing underlying constructs. Messick not only questioned the unidimensionality of SDR (assumed by Edwards' and Crowne and Marlowe's measures), but also put in doubt that social desirability can be measured in many research populations, as subjects may differ markedly in their "conformity to the group consensus of desirability". It is also worth to note, that Messick had significant problems in interpreting the isolated factors. The research with the use of factor analysis as a main tool continued and brought interesting results in the work of Wiggins (1964), who factor-analysed many scales, including the MMPI, Edwards' SDS, MCSDS and various tools measuring authoritarianism, acquiescence and related concepts. The analysis yielded six factors, three of which were interpreted by Wiggins as related to SDR (one was related to acquiescence, one to conformity and one to construct related to authoritarian personality). Among the three factors related to SDR Wiggins differentiated "non-endorsement of pathology" (labelled alpha or "popular responses"), "endorsement of desirable but infrequently possessed traits" (labelled gamma or simply "lying")[8] and "cautious, controlled good-impression". The last two scales were

[8] The little informative labels „alpha" and "gamma" are ascribed to Block (1962/1965), who suggested these neutral terms to describe the three MMPI stylistic dimensions he discovered due to huge controversies regarding their interpretation. Factor I was labelled "alpha", factor II "beta" and factor III "gamma". Factor alpha was thought-of as self-deception, beta as acquiescence and gamma as lying/faking. Block's work was published in

correlated with each other. The general pattern was similar to the one obtained by other researchers in a similar time, e.g. Edwards, Diers and Walker (1962). This evidence also disconfirmed the unidimensionality of socially desirable responding, but proposed only tentative approaches regarding the interpretation of the isolated factors. Wiggins also noticed that some of the newly proposed one-dimensional SDR scales, including the MCSDS, loaded on more than one SDR factor, among those found in his research (1964).

2.1.5 First integrations of the field

Damarin and Messick (1965) were the first to provide a comprehensive review of response styles research to date, they have also proposed more elaborative theories about the discerned communalities between the SDR scales. To perform this analysis they have reviewed more than 20 factor analytic studies based on questionnaires and performance tests that were conducted on different samples (e.g. college students, school children, air force cadets, etc.). Data from many laboratories were used and all major response biases then-known were the scope of their research (acquiescence, social desirability responding, extremity response style).

The researchers noticed that the desirable responding "seems to be at once the most important and the most elusive" of all the response styles, pointing to many difficulties in measuring it. Among many others they pointed to the lack of methods to verify participants' accuracy, which is needed both to perform criterion validity studies and to truly verify that someone is misrepresenting one's personality or abilities ratings. Moreover, they observed that many measures of SDR tendencies were in fact convolution of many variance sources, disentangling of which was not always possible using the factor analytic methods. Finally, they have emphasised the importance of theoretical advances that were needed to interpret the growing and confusing body of evidence.

Their extensive review led to the differentiation of two main factors underlying the SDR concept- one of them related to bias in self-regard and the other linked with bias in self-report (1965). The former is believed to stem from an interaction of certain traits and attitudes towards the self with the process of responding in self-reports. Damarin and Messick enumerated various related constructs like anxiety, self-esteem or self-confidence. They also turned attention towards autism[9], which they defined as affective distortion of the "cognitive picture" (attitudes, judgments, etc.) leading to misconceptions about one's traits and attitudes. This "autistic bias of self-regard" (Paulhus, 1986) was described as maintaining one's positive image by distorting the reality in favour of the self once the self-image was threatened. In this sense this factor was related to the self-deceptive and defensiveness line of the SDR research and was linked with ego-protection/resiliency concept. Damarin and Messick linked it also to the research of Frenkel-Brunswick (1939), psychodynamic theories of repression and also to some very much practical questions of the accuracy of self-appraisal. They have also pointed out that the discovered multidimensionality of these tendencies (e.g. Messick, 1960) might be related to the fact that what was perceived as socially desirable or undesirable seemed to be characterised by substantial inter-group and inter-individual variability. This remark was used as a basis for a call of more studies looking for individual differences in response biases.

1965 as a book, but circulated as a manuscript already from 1962, hence the terms are attributed to him despite the Wiggins' article from 1964 was probably the first to use them in (official) print.

[9] Damarin and Messick (1965) use this term according to Murphy (1947) who defined autism as a state when "affective states distort the cognitive picture of reality so that the person is misled". This definition and its implications is of course very distant from the currently used meaning of the word "autism" in, e.g., clinical psychology.

On the other hand, the bias in self-report was called as “propagandistic bias”, “aimed at producing a specific effect on a specific audience”. This bias was hence regarded as a motivated, deliberate manipulating of one’s image in order to achieve certain “goals and purposes”. It was stated that the content of such self-presentations could vary from occasion to occasion according to what image was deemed more socially approved on a given occasion. This bias was related with the need of social approval and conformist behaviour. However, using the MCSDS to measure it was not advocated, as the scale correlated with both biases which pointed to its interpretation as a blend of various constructs. The propagandistic bias was related to the faking/lying line of research, withal Damarin and Messick criticised the use of such umbrella, sobriquet terms as “faking” or “lying” as, in their opinion, they did not reflect the level of precision needed to describe all the relevant stages “of awareness” between “conscious accuracy and conscious misinterpretation in self-report” (1965).

The researchers also pondered on the issue of self-report accuracy and its relation with SDR. The point of departure of this thinking was a constatation that self-reports are distorted also in low-stakes, non-threatening and anonymous research conditions where, in fact, there is no special pressure on participants to respond desirably. Hence, they have put forward a theory that the observed self-report score is an interplay of SDR tendencies, desirability of characteristics underlying the items, accuracy of self-perception and biases in the accuracy of self-perception and in the perceived desirability of characteristics. Obviously, to assess an individual’s self-accuracy some external criteria (non-self-report measures) were needed. Regrettably, the researchers did not develop this theory further on (Messick, 1987), despite that some other researchers confirmed the importance of self-accuracy in reporting (e.g. Paulhus, 1984).

Damarin and Messick (1965) also noted that the two identified dimensions, bias in self-regard and bias in self-report, were “relatively uncorrelated”. This was interpreted as a piece of evidence to corroborate the dual nature of SDR- as a personality-like trait and as a situational phenomenon. This view was present in the field since the very first research by Hartshorne and May (1930). The Damarin and Messick’s research also concluded that bias in self-regard was responsible for more variability in the questionnaire scores than bias in self-report.

The deepened analyses of SDR and its factors brought by Damarin and Messick surprisingly failed to fuel further inquiry in the field as the SDR research reached its peak in 1960s and then saw a decline in the subsequent decade (Paulhus, 1986; 1991). Probably this was caused by a certain inertia of the methods used- the knowledge could not be pushed further with simple factor analyses of self-reports. It needed more theoretical and methodological advancements to move forward again.

2.1.6 Conception of Sackeim and Gur

The subsequent research was continued by Sackeim and Gur (1978) who concentrated their efforts on the larger of the two SDR factors: bias in self-regard/self-deception. The researchers concocted experimental psychology, psychophysiological methods, philosophy of mind and psychodynamics theory in a mixture that managed to bring new point of view, much needed in the response biases field. Sackeim and Gur (1978) departed with a question: “How it is possible at all, that people are capable of self-deception?” They continued with a statement that self-deception would constitute an insolvable paradox if the consciousness would be regarded as unitary and transparent. Hence, no serious research on self-deception is possible without assuming that consciousness is at least non-transparent, namely that an individual does not have full and immediate access to every element (e.g. thought, judgment, etc.) of consciousness. Sackeim and Gur also argued that this non-transparency of consciousness is actively organised and motivated by an individual, in other words, that there is a certain goal behind each process of self-deception. The authors saw this goal in a defence against

threatening stimuli and perceived self-deception as a grounding for all defensive mechanisms. Importantly, they have also added that "motivational" does not mean "intentional", which differentiates self-deception from lying or faking and argued that unintentional lying is not possible, whereas unintentional deception (self- or other-) is. They have also formulated four conditions necessary and sufficient to speak about self-deception: "a) The individual holds two contradictory beliefs (*p* and not-*p*). b) These two contradictory beliefs are held simultaneously. c) The individual is not aware of holding one of the beliefs. d) The act that determines which belief is and which is not subject to awareness is a motivated act." From these points it can be clearly seen that their theory was heavily influenced by a psychodynamic approach. In other words, Sackeim and Gur proposed that a given individual at the same time holds two contradictory judgements: e.g. "I am good at math" and "I am not good at math", but that only one of these beliefs is available to the conscious self. The non-threatening "I am good at math" will be selected as the negative "I am not good at math" would be most probably frowned upon by teachers and/or parents and would be threatening to one's self-esteem.

Sackeim and Gur (1978) brought some evidence to back-up their theory. In example, in an experiment where participants had to recognise voices on tape recordings some of them did not recognise their own voice while their psychophysiological responses (e.g. skin conductance) showed that they have recognised it correctly. Moreover, the rate of these denials correlated positively with self-report measures of self- and other-deception constructed by the authors (1978). In an another article Sackeim and Gur (1979) presented validity evidence for their two 20-item each measurement tools: Self-Deception Questionnaire (SDQ) and Other-Deception Questionnaire (ODQ). The SDQ items concerned mainly threatening and unpleasant thoughts and feelings: hatred, guilt, aggression, socially disapproved sexual fantasies, etc., in general "statements (…) universally true but psychologically threatening" (Paulhus, 1984) and with a "psychoanalytic flavour" (Paulhus, 1986). The items had an entirely intrapsychic character as they all treated about internal psychic states and feelings unobservable by anyone from outside. On the other hand, the ODQ was a conglomerate of items about overt behaviours that were "culled from various lie scales" (Sackeim & Gur, 1979) and depicted socially approved but practically infrequent behaviours. The scales correlated negatively with measures of psychopathology, e.g. depression. The authors concluded that this did not mean that self-deceivers are on average absolved from psychopathology but that they are less susceptible to admit to it. They also saw this result as yet another sign that self-deception is a greater threat to self-reports validity than other-deception. The SDQ and ODQ correlated positively in their research, suggesting that self- and other-deception might have more in common than assumed by Damarin and Messick (1965; Sackeim & Gur, 1979).

2.1.7 Paulhus' models: toward modern conception of SDR

Paulhus (1984) stepped in next to verify the SDQ/ODQ validity and to deepen the research on these two promising scales. He tested their relationship to the Alpha/Gamma model proposed by Block (1965) and Wiggins (1964) by conducting the convergent validity study in which he included all the most popular SDR measures then available: MMPI K scale ("lie scale"), Edwards' SDS, MCSDS and Wiggins' SD scale. The factors that emerged from the analysis clearly resembled the Alpha/Gamma dimensions and supported the interpretation of SDQ as a measure of self-deception and the understanding of ODQ as a gauge of other-deception. However, the precise analysis of the content of the items that loaded on the other-deception factor pushed Paulhus to slightly alter its name and interpretation (1984).

Moreover, he noticed certain psychometric and methodological faults of the SDQ/ODQ. The most important of them were: all the SDQ items were negatively-keyed, whereas the ODQ only had

positively-keyed positions. This caused obvious methodological problems but also interpretational complications- some might rise a suggestion that the SDQ did not measure self-deception but rather denial of negative attributes and that the ODQ was not a measure of other-deception but of accepting positive attributes. This posed a clear threat of confusing the traits of interest with acquiescence response style. What is more, not all of the items loaded highly on their destined factors and some of them had minor wording problems. Paulhus (1984) managed to fix these problems by rewriting the items and balancing the keying within the scales. He also conducted another convergent validity study with the new items and confirmed its construct validity by obtaining correlations with other SDR measures consistent to the theory and by extracting the two predicted factors. Similarly as in the Sackeim and Gur scales (1979) the two factors were correlated, as confirmed in a subsequent confirmatory factor analysis where an oblique model fitted the data better than an orthogonal one. Paulhus named the factors self-deception (SD) and impression management (IM) and called his scale Balanced Inventory of Desired Responding (BIDR). This scale generated significant interest from the researchers worldwide but never completely replaced the MCSDS as a main SDR-measuring tool despite its obvious supreme theoretical and psychometric characteristics[10].

In the following endeavours Paulhus (1984) conducted an experimental study in which he used an instruction manipulation paradigm similar to the one devised by Ruch (1942). The experiment used a between-subjects design where one group answered the BIDR items under standard research instructions, in a large-group setting and anonymously. Second group answered them in small groups in a proctored session and had to give their name and phone number. Moreover, the instructions stated that the experimenters (proctors) were going to review their answers to check if they had understood the questions correctly. In this manner Paulhus simulated public condition in which he expected the impression management tendencies to emerge more than in anonymous condition. The results obtained confirmed his expectations: the IM scores were higher under public condition than under anonymous condition but the self-deception scores remained the same in two conditions. This urged Paulhus to constate that impression management is a larger threat to the validity of self-reports than self-deception (this proposition was contrary to the main beliefs in the field, cf. Sackeim & Gur, 1978). He also advocated purging its variance from any measures as in his opinion IM was only a response set, entirely context-dependent and without any substantive relations to psychological traits ("any intrinsic relation to central content dimensions").

Paulhus also recommended further research on self-deception character which he believed was comprised of two underlying dimensions: defensiveness against psychologically threatening motives and an as yet unidentified trait related to high self-esteem and low anxiety. He concluded that the latter dimension was a convolution of self-esteem, SDR, anxiety and accurate self-reports that was "difficult to tease apart psychometrically" (1984). Paulhus did not precise what motive would underlie this third dimension, however, in his next text (1986) he reflected on disentangling this type of self-deception from a presentation of a genuinely well-adjusted, high self-esteem participant. He concluded that a moderate self-deception may have adaptive value and may help in coping with everyday difficulties. Paulhus also stated that this mild self-deception may underlie a number of psychological constructs, including control perception, achievement motivation and social dominance (1986). He also mentioned that due to the lack of adequate methods it was unable to tell apart self-deceivers from participants accurately yielding a positive image of themselves. Paulhus concluded that

[10] It would probably take another PhD to precisely calculate the interest in both scales along the years but the Marlowe and Crowne's main publication about the scale has around 5500 citations so far, including 80 in the year 2019, whereas the Paulhus' 1984 article was as yet cited only 3500 times of which almost 200 come from the year 2019. All the numbers come from Google Scholar and obviously can be treated only as very unrefined measures of the scales' popularity.

probably all self-report measures of self-deception were to some extent a compound tapping adjustment, self-deception and other, unmotivated biases at the same time (1986). He also came forward with a proposition to differentiate between acute and chronic self-deception, the first being a short and transient form of this phenomenon. The unmotivated biases were considered distinct from SDR as driven by cognitive or informational biases (Nisbett & Ross, 1980). Cognitive biases stem from the way the human cognitive system process information and include, among others, memory biases, e.g. hindsight bias (Campbell & Tesser, 1983). Informational biases, on the other hand, are derived from restrictions on information availability, e.g. on the basis of social conventions that in general enjoin refraining from negative remarks about someone's performance, behaviour, looks or personality or that admonish for passing bad news to people (cf., Tesser & Rosen, 1974). Hence, the research on self-deception needed new methods and new theories in order to unravel the self-deception riddle. However, it was concluded that probably the positive bias of self-perception is a universal tendency that manifests itself in many (all?) traits (Paulhus, 1986).

Paulhus also integrated new evidence on impression management (IM), concluding that there were three possible views of this construct: a) strategic presentation, b) motive, c) skill (1986). The first one claims that every social situation has an ideal image that can be presented to "the audience" in order to achieve some instrumental gains. Jones and Pittman (1982) presented a taxonomy of IM strategies, that included ingratiation, threatening and self-promoting among others. However, as it was pointed by DeMaio (1984) in many research situations "the audience"[11], especially defined as the society in general, is a very vague concept and it seems unlikely that in such a wide gamut of measurement occasions participants would engage in IM only to pursue "instrumental gains". In fact, in most of the research situations there is nothing to gain, even if we include social reinforcements here. This view was supported by Paulhus (1986; 1991) who claimed that in most of the cases IM is not a presentation in order to gain something, but a purpose in itself. In this sense IM is an autotelic motive to be liked by others or to achieve a respectful social position. Similarly, self-deception can be seen as an autotelic motive to like yourself. Finally, IM can be also viewed as a skill- some participants may be able to manipulate their images whereas other can be characterised by poor abilities in that matter. It was also deemed probable that certain personality types resort to IM more often than others (Paulhus, 1986).

Subsequent works were concentrated on refinement of the theoretical classification of different SDR types. Moreover, alternative explanations to those formed by Paulhus (1986) were formulated, as some researchers, e.g. Roth, Snyder and Pace (1986), proposed that the differentiation on self-deception and impression management is not necessarily correct as the two-factor structure of SDR can be explained by alternative dimensions: enhancement and denial. The former was simply defined as affirming positive items (e.g. *I am good at math; I am saint*), the latter was denying negative items (e.g. *I am bad at math; I am a sinner).* However, the scale prepared by Roth et al. (1986) did not separate completely keying direction from positivity of an item, hence a certain ambiguity was left regarding whether which model was actually confirmed by their study.

A thorough comparison of the two models was done by Paulhus and Reid (1991) who wrote 20 additional items to the BIDR scale in order to measure self-deceptive enhancement and denial tendencies separately. The resulting questionnaire, named Self-Deceptive Denial scale (SDD) , was similar to the SDQ by Sackeim and Gur (1978) and contained mainly items denying aggressive behaviour and sexual fantasies (e.g. *I can't think of anyone I hate deeply; I have never felt like I wanted to kill somebody*). Paulhus and Reid (1991) conducted three correlational studies where this new scale

[11] This notion appeared in the seminal papers of Damarin & Messick (1965) and Sackeim & Gur (1978), cf. Paulhus & Reid (1991).

was used along with other questionnaires: SDR, self-esteem, self-monitoring and other measures. The results showed that only the self-deception factor can be divided into enhancement and denial types, the impression management factor did not form separate dimensions. Moreover, self-deceptive denial subtype correlated to a large extent with the impression management scale, while correlated only modestly with self-deceptive enhancement (Paulhus & Reid, 1991; Study 1 and 2).

In order to interpret the substantial interpretation of the emerging types of SDR correctly and to disentangle the keying-valence confound Paulhus and Reid (1991) constructed yet another scale where they separated keying direction from item valence. As a result a 2x2 design was created with items like *I am a saint* and *I am not a sinner* differing in valence but not in keying direction. A socially desirable response is "yes" for both the former and the latter while the former contains a positive word "saint" and the latter a negative one- "sinner". On the other hand, positions like *I am a sinner* and *I am not a sinner* differ in keying direction but not in valence, as a socially desirable response is "no" for the former but "yes" for the latter, whereas both items contain a negative word "sinner". The results of the study yielded much higher correlations between the items that have the same valence and different keying than *vice versa*, e.g. items *I am a sinner* and *I am not a sinner* correlated higher than items *I am a saint* and *I am not a sinner*. Paulhus and Reid interpreted them as pointing in favour of the enhancement/denial (valence) hypothesis over the keying direction one (1991, Study 3). Hence, the two new types of self-deception were called self-deceptive enhancement (SDE) and self-deceptive denial (SDD). What is more, they have also provided a set of intercorrelations between SDR types and other measures that showed that SDE is related to good adjustment, e.g. high self-esteem and low anxiety, indicating that self-enhancing is somewhat related to building positive esteem. Paulhus and Reid hypothesised that the mechanism of it may be creating an illusion of control and competence and/or rejecting negative feedback (1991). SDD and IM are related closely to each other, which can be interpreted that denial behaviours are related more to conscious, other-related deception than unconscious self-deceiving. Paulhus and Reid also saw a possibility that the sensitive character of the SDD items was responsible for triggering such behaviour (1991).

Nonetheless, the research of Roth et al. (1986) and Paulhus and Reid (1991) resulted in conclusion that the self-deception variance could be separated into two distinctive constructs: self-deceptive enhancement and self-deceptive denial. Both were believed to be distortive for self-perception but not related to the presence in front of an audience. It is also worthy to note that these conclusions meant that with every other further refinement more and more SDR types were proposed as the theory evolved from a one-dimensional proposals (e.g. Edwards' conception) to a three-dimensional one (Paulhus and Reid, 1991).

The appearance of new SDR forms led to an attempt to study their nomological network further on. This endeavour was taken up by Paulhus and John (1998) who analysed links between SDR types and numerous socio-psychological traits. In their analysis of the relation between SDR and personality traits (Big Five framework was used) they found out that both self-deceptive enhancement and self-deceptive denial correlated with personality traits but that they also formed a distinctive network, similar to a two dimensional structure of self-deception: enhancement was related to traits like surgence, dominance or intellect, whereas denial correlated with agreeableness, nurturance and dutifulness (similar results in Paulhus et al., 2003, where enhancement was related mostly to extraversion). According to the expectations, self-enhancement also correlated with narcissism[12], as narcissists commonly exaggerate their achievements and skills (Raskin & Hall, 1981). Paulhus and John (1998) aimed to interpret this pattern by referring to two values and motives identified in sociology

[12] In this line of research narcissism is treated as a personality trait, not as a clinical syndrome.

and social psychology: agency and communion. These two notions were identified as basic values introduced to everyone in the process of socialisation. Agency is understood as valuing strength, competence, independence, individuality, achievement, whereas communion means placing value on taking care about relationships, being a good, rule-abiding, agreeable member of society (Bakan, 1966; Paulhus & Trapnell, 2008; Wojciszke, 1994). Communion and agency were also defined as "general thematic clusterings (…) which may be mirrored in conscious values, specific attitudes, particular interests, stylistic traits, characteristic self-schemata and social motives (…)" (McAdams, 1985). Paulhus and John (1998) concluded that the interpretation of self-deception in the light of agency-communion framework would mean that self-deception is much more than just a response style, but it should be rather understood as a fundamental trait of human self-perception as the two basic values "give rise to two motives: need for power and need for approval". These motives result in inter-individually varied tendencies towards exaggeration of agentic or communal traits. This exaggeration results in systematic biases of self-perception. According to this model "values lead inexorably to biases" (Paulhus & John, 1998). Importantly, both tendencies were ascribed adaptive, self-protective character. Moreover, it was suggested that men and women should have different proneness towards the two biases: men should incline more into agentic bias (deception on traits related to agency), whereas women should yield more moralistic bias (deception on traits related to communion) (Paulhus & John, 1998).

These results brought change in the theoretical SDR models proposed by Paulhus (1986). In the earlier models bias consciousness was the main difference between the SDR types: self-deception was believed to be an unconscious tendency and impression management- a conscious one. The new results prompted an amendment to this model by setting agency *versus* communion as the new critical difference between different SDR types. Within the two main frameworks both conscious (impression management, faking) and unconscious (self-deception, enhancement, denial) biases are possible (Paulhus & John, 1998). In line of these conclusions, Alpha and Gamma factors proposed by Wiggins (1964) and Sackeim and Gur (1978) were reinterpreted as agentic and communal bias, respectively. In this way the noncommittal labels were put into an interpretative framework and set into a nomological network of antecedent motives and correlated traits.

This understanding of SDR would explain its prevalence even in low-stakes, low-threat contexts. According to this line of reasoning SDR, strictly speaking one of its types, self-deception, is always present in every measurement situation that require any self-insight or self-evaluative action from a respondent as any of these actions is distorted by inherent biases of self-perception. These biases led to an over-optimistic, over-positive opinion about oneself, an exaggerated "global self-positivity" (Paulhus & John, 1998). It is worthy to note, that the scope of self-deceptive biases is much larger than previously postulated by Sackeim and Gur who limited them to only sexual and aggressive (in general: socially inappropriate) content (1978). Moreover, an adaptive role of self-deceptive tendencies and their inter-individual variability were also proposed. What is more, the results and interpretations of Paulhus and John (1998) for the first time brought an idea that SDR can actually be both a response style (in case of self-deceptive tendencies) and a response set (in case of impression management), depending on person characteristics and/or measurement context.

2.2 Modern views on SDR

2.2.1 Paulhus' (2002) model

In his seminal article from 2002 Paulhus summed up the history of the SDR research and briefed the state-of-the-art. SDR was defined as an overly positive image of oneself, with the word "overly" being a critical word in this definition as it is the departure from reality that is pivotal to classify a desirable,

positive self-description as a response bias. Another crucial consideration regards SDR character and context: both proved to be much more general and widespread than it was previously believed. SDR evolved from a concept present in a limited number of occasions, mainly threatening, intrusive situations (cf. Sackeim & Gur, 1978; 1979) and episodes of seeking social approval (Marlowe & Crowne, 1960), to a general phenomenon present in every instance when self-report is performed. Moreover, the previously accepted key differentiation between SDR types that concentrated on the degree of awareness (e.g. Paulhus, 1986; Sackeim & Gur, 1978) has been changed to a crossed classification where consciousness of bias and content of self-report is equally important. Paulhus and John (1998) pointed that content can define the type of self-presentational strategies. This landmark conclusion helped to clarify certain issues in the SDR theory, e.g. why self-deception denial items (SDD) fell on the same factor as impression management items (IM)- simply they were both communal in character. Moreover, this new classification enabled to show that SDE and SDD scales, not only IM ones were also sensitive to instruction manipulations, just the content of instructions has to prompt agentic enhancement, not only communal motives. The most up-to-date SDR classification proposed by Paulhus (2002) contained four bias types (presented below).

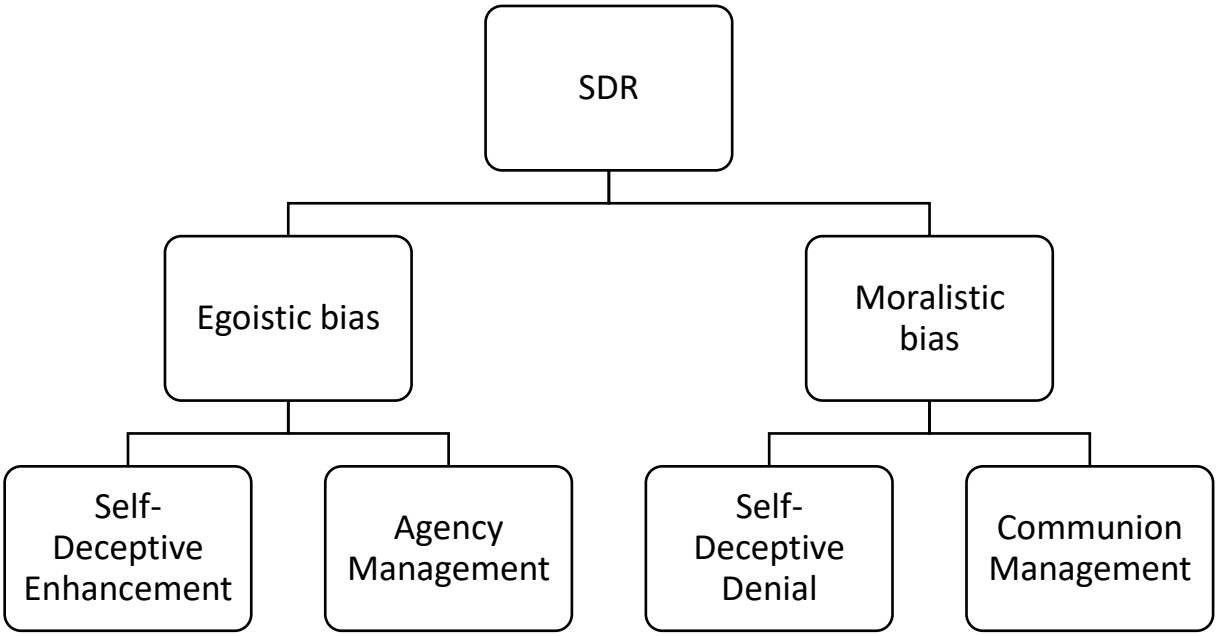

Figure 1. SDR classification system proposed by Paulhus (2002). An extension of the Paulhus (1984) model where SDR was divided only on self-deceptive enhancement and impression management.

This new structure of SDR was reflected by Blasberg, Rogers and Paulhus (2013) who proposed a refinement to the BIDR measurement tool in order to contain the most recent theoretical advancements. New scales were proposed to measure both agentic and communal impression management, called jointly Bidimensional Impression Management Index (BIMI). The authors showed validity evidence that enables to confirm the BIMI addition to the BIDR as a valuable tool to measure impression management tendencies. However, to date the BIMI scales did not enjoy much interest from the researchers and its value is largely unverified.

The acceptance of the above classification, linked closely to the basic meta-values of agency and communion framework, also caused a diversion from seeing impression management as simply blatant

lying (as in e.g. Eysenck & Eysenck, 1968; Hathaway & McKinley, 1951; c.f. "lie scales" research). The new conception enclosed IM to self-presentation framework and the deliberate, motivated dissembling was practically limited only to a handful of contexts, among which job application occasions predominate where IM is most often called "faking" (faking good). Other context where a goal-oriented IM takes place is the forensic and clinical setting where both faking good and faking bad (malingering) may take place as a way to distort measurement tools in order to get a benefit (e.g. avoid punishment, get out of jail, obtain health leave). Interestingly, recent evidence suggested that faking good and faking bad may be driven by distinct mechanisms and that they are not just two sides of the same coin (Bensch, Maaß, Greiff, Horstmann, & Ziegler, 2019). Nonetheless, the current dissertation is focused on the other branch of the SDR tree and faking, either good or bad, is not in the centre of attention here. Excellent book-length compilations on faking (good) were edited by Griffith and Peterson (2008) and Ziegler, McCann and Roberts (2011). Malingering was covered e.g. by Morey and Lanier (1998) and Rogers and Bender (2018), see also Bensch and others (2019).

Basing SDR hierarchy on the agency-communion framework leads also to interesting questions about the relations between distinct forms of SDR with each other. Results presented by Blasberg et al. (2013) showed that agentic and communal IM scales correlated[13] only 0.10 under honest conditions, but 0.44 under fake good conditions. This correlation was believed to stem mainly from "non-faking cluster" (around 11% of sample did not fake and this caused a low cluster-low cluster relation that induced inter-variable correlations). Interestingly, both scales were reactive to fake good instructions-researchers concluded that both agency and communion are valued and that the shape of their relation depends on the measurement context (Blasberg et al., 2013). This pattern of relationship is corroborated by Wiggins (1991) who determined the two dimensions as separate but with a possibility to interact or even enhance each other in certain social contexts. Wiggins claimed that "all combinations are possible", hence an individual can yield large agentic bias and no communal bias or *vice versa*, no bias in both dimensions or large in both as there is no conflict between the dimensions (1991).

The new classification of SDR also entails that under this term a response set or a response style can be meant thus ending the long dispute whether SDR is a set or a style. However, the acceptance of this view sparked a new debate concentrated around the question whether SDR is just a style or whether it can be thought of as a substance, namely a personality trait. Hence, the set versus style dispute was replaced by, equally heated, substance versus style debate. Some researchers noticed that positive, socially desirable self-reports are often confirmed by external sources of information, e.g. peer reports (McCrae & Costa, 1983). Moreover, partialling out SDR variance measured by SDR scales many times led to lowering the external validity of self-reports, thus indicating that the scales capture also (predominantly?) substantive, not spurious variance (Birenbaum & Montag, 1989; McCrae & Costa, 1983; Ones, Viswesvaran & Reiss, 1996). There were also conceptions that SDR is in fact a personality trait, e.g. Block proposed ego-control and ego-resiliency traits that could capture the SDR variance (1965). There were also voices that SDR is not a serious threat as participants do not distort their self-reports to a significant degree (Ones et al., 1996). Many of the arguments were refuted by Paulhus (2002) but the substance *versus* style discussion will be revisited in the section devoted to SDR control methods (4.5).

[13] Whenever a specific value of a "correlation" is given it is meant as a value of a zero-order (unpartialled, raw, gross) Pearson correlation coefficient unless explicitly stated otherwise.

2.2.2 Further integration: nomological network

Paulhus also summed up the research on SDR nomological network that was significantly broaden by modern research efforts (2002). Personality correlates of SDR were found as the SDR scales correlated negatively with emotional stability and openness to experience and positively with extraversion and conscientiousness (McCrae & Costa, 1983; Ones et al., 1996). Surprisingly, agentic (Alpha-type) bias was related to intelligence measures (IQ). Also, larger agentic management for male than female and for Europeans than Asians was identified, pointing to cultural differences influencing SDR tendencies (Blasberg et al., 2013). Moreover, Paulhus and John (1998) suggested that "low cognitive complexity" participants should yield more SDR bias. Paulhus and Trapnell (2008) reported IM links with SDE, narcissism and hindsight bias, confirming earlier results of Paulhus and John (1998).

Similar correlations with self-esteem, life satisfaction and depression scales led researchers to conclude that SDR can be related to good adjustment and have beneficial consequences for an individual (Paulhus & John, 1998). It is worthy to note, that only suggestions for self-enhancement (SDE) with adjustment were raised, impression management and self-deception denial seem unrelated to it. However, recent meta-analysis showed that benefits stemming from self-enhancement are indeed positive for personal adjustment, but they are "a mixed blessing" for interpersonal adjustment: most of the benefits in this area are most likely short-lived and self-enhancement can even lead to long-term deterioration of interpersonal adjustment (Dufner et al., 2019). Complementary, but more nuanced evidence, was presented by Taylor and Armor (1996) who suggested that self-enhancement as a trait may not be adaptive, but when used only as a strategic behaviour, applied when needed to boost self-esteem, it may lead to positive consequences. It is worthy to note the notions of "unmitigated agency" and "unmitigated communion", both stemming from Bakan's theory (1966). The two define extreme actions where one dimension of the dichotomy is promoted regardless the other and both can result in adverse inter- and intra-personal consequences (Helgeson & Fritz, 1999; Wiggins, 1991). It is probable, that negative long-term consequences of self-enhancement may come from unmitigated manifestations of agentic motives.

All this key SDR discussions, substance vs. style and adaptive vs. maladaptive, made researchers aware that new methods of measuring and controlling SDR are needed. Especially the need of criterion-related measures was voiced. Paulhus and John (1998) popularised new technique named self-criterion residual (SCR) of comparing self- and other-reports treating the latter as a criterion of self-reports. Moreover, new SDR scales, such as the BIMI, were proposed.

Apart from methodological refinements, the SDR framework needs also further theoretical works including efforts to merge this paradigm with many other conceptually-related areas from the domain of social sciences. Paulhus and Trapnell (2008) initiated a series of attempts to integrate SDR research with related research fields of self-presentation, self-monitoring, self-enhancement, Dark Triad personality and self-esteem. Merging theoretical frameworks of related concepts enabled Paulhus and Trapnell (2008) to distinguish three trait-like characteristics related to individual differences in overly positive self-reports: a) attunement to self-presentational demands (attention to self-presentation), b) motivation to engage in self-presentation, c) amount of distortion in self-presentation (nature of image intended to present). The first trait would describe attending to demands of a given social situation and tailoring one's behaviour to them. Self-monitoring and self-consciousness are among the traits that are thought to describe such social attunement. Both are often treated as skills that enable to gauge social demands and act according to them. Self-consciousness was often divided further into private, public and social anxiety (Buss, 1980). Other approach sees social attunement as a consequence (a facet?) of the classical personality trait- extraversion (John, Cheek & Klohnen, 1996).

Nonetheless, the Snyder's approach of self-monitoring won the biggest popularity, probably also thanks to the research instrument conceived along with the theory (1974).

Numerous motives were enlisted as possibly leading to self-presentational behaviours, among the most important frameworks were: self-promotion/self-protection (Arkin, 1981); self-enhancement, self-verification, self-evaluation (Swann, 1990; Leary, 2007); egoist (self-enhancement), politician (popularity), consistency-seeker (consistency) and scientist (truth) typologies (Robins & John, 1997). Finally, the nature of image presented is based on agency and community dimensions. However, there are difficulties with finding trait-like characteristic that would be related with the degree of distortion. Despite this research lacuna, it was sufficiently evidenced that the two fundamental meta-values organise the content of self-presentation and constitute two basic types of this behaviour (Lonnqvist, Verkasalo & Bezmenova, 2007; Paulhus & Trapnell, 2008).

2.2.3 Paulhus and Trapnell's (2008) model

Paulhus and Trapnell (2008) proposed to reorganise SDR theory along the two fundamental dimensions: the first one being context, the second content. The former is understood as contextual demands on self-presentation as exerted by measurement situation. The most important characteristic here is audience- if it is present then the context is deemed as public and it is related to higher SDR motivation. If it is not present, then the context is called private as the only "audience" is the self. Another important characteristic of measurement situation is stake, with high-stakes situations eliciting SDR more easily. Content is considered within the agency-community framework- different image of the self may be evoked in different social situations, also contents may inter-individually differ in subjective importance (desirability).

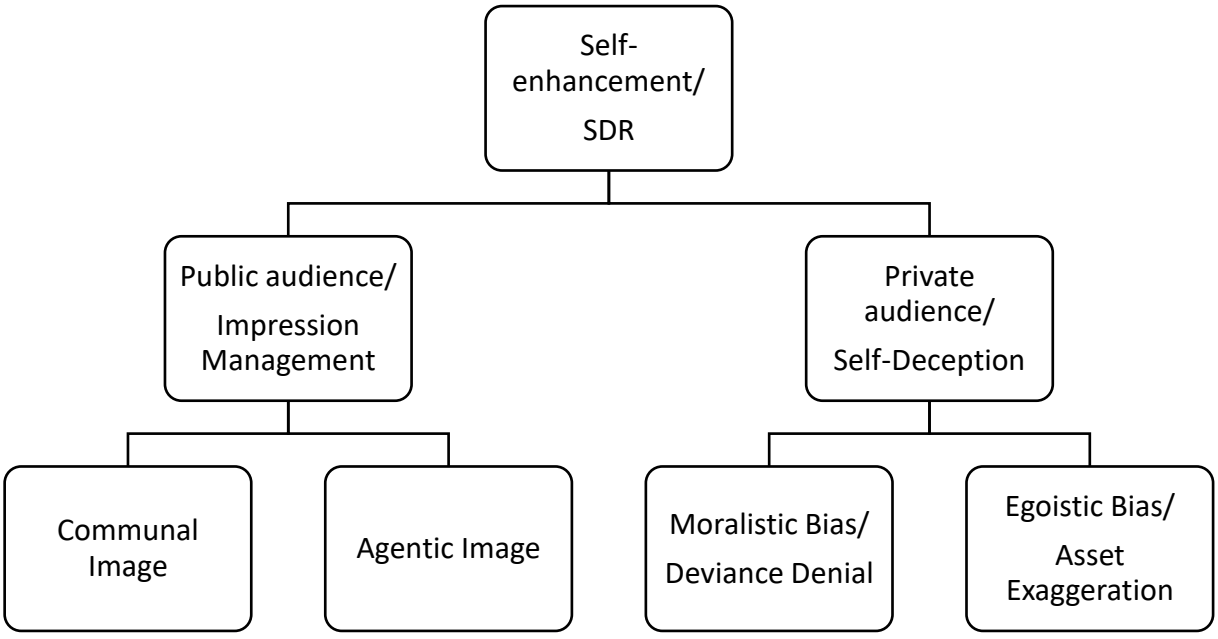

Figure 2. Schematic organisation of self-enhancement and SDR as in the conception of Paulhus and Trapnell (2008). The concept divides SDR/S-E because of context (audience) and content (agency and communion).

Such organisation of distortive responding also enabled to untangle some of the long-debated controversies. In example, Paulhus and Trapnell (2008) contended that SDR is neither response set, nor response style, nor accurate report of a desirable personality, but can be either of them depending on the situation (cf. Paulhus, 2017). An *ad hoc* presentational strategy due to situational demands would be called a response set, whereas a stable individual tendency to yield distorted self-evaluations would be treated as a response style. By adopting the hierarchy of image content and audience presence, Paulhus and Trapnell (2008) were also able to circumvent the unfeasible problem of consciousness role in SDR classification. Now consciousness is not a key characteristic used to classify response distortions and is not needed to predict nor explain them (see also Paulhus, 2002; Pauls & Crost, 2004). These new organisation of concepts enables also to explain self-presentational behaviours in low-stakes, private and anonymous situations. These instances are now often classified as automatisation of self-presentation, habitual self-presentation or an effect of other social motivations, e.g. self-enhancement.

Most importantly, Paulhus and Trapnell (2008) pointed to a conceptual overlap between many dispersed terms that have common content under "disparate labels". Among these terms are: socially desirable responding, self-enhancement, self-presentation, impression management- differences between them are of mainly historical nature, as origin of studies on them was different. Nowadays, however, it is time to merge them, which is possible under the classification proposed by Paulhus (2002) and Paulhus and Trapnell (2008). Their most important common core from the point of view of social sciences methodology is the effect to which they lead- a distorted, overly positive bias in self-descriptions.

Hence, different overly positive biases were integrated under the two-level framework with context and content as organising factors. However, there is still a great need to further research phenomena underlying the observed effects. Apart from the bridge towards underpinning psychological processes there is also a great need of new research methods to measure and control the overly positive biases (Paulhus & Trapnell, 2008; Ziegler, 2011; 2015).

2.2.4 Positivity bias emergence

Since Paulhus' theoretical advancements (2002) no systematic review of SDR was published (Bensch, 2018), but some of the following results showed that further differentiation between the phenomena and building the nomological network of SDR, faking and related concepts is much needed (Bensch et al., 2017; 2019). Such endeavours were undertaken by Matthias Ziegler, and a large group of his collaborators, and led to substantial advancements in SDR knowledge.

First of all, they have proposed to put numerous but ill-defined concepts into one nomological network and conceptualise them under one term of **positivity bias** (Anderson, Brion, Moore & Kennedy, 2012; Bensch et al., 2017). Furthermore, a merging effort was expanded also to the conceptions of spurious measurement error (Schmidt et al., 2003) and method variance (Podsakoff et al., 2003) pointing to the positivity bias as a source of systematic measurement error that leads to increase in spurious variance in measurements. Such a conceptualisation enables to treat response biases as sources of variance and account for them in statistical models. Moreover, Ziegler (2015) pointed that also other response biases should be studied together with positivity bias, including response styles, as some initial research revealed unexpected cognation between MRS and SDR in qualitative research (think-aloud verbal protocols) which is a good incentive for future studies (Ziegler, 2011). Most importantly, Ziegler (2011; 2015) complemented models of cognitive survey responding (Krosnick, 1999; Tourengeau, Rips & Rasinski, 2000) by adding person (trait) and situation (context) characteristics that may interact with each other and can exert their influence on every stage of survey responding. This account also entails

that positivity bias is domain-specific, namely not every item is biased and not every biased item is distorted to the same degree. Moreover, this bias does not always occur but occurs always when given circumstances take place. Its occurrence is believed to be a function of interaction between person and situation characteristics. Among key personal characteristics Ziegler (2011) named e.g. general mental ability, self-reflection, honesty and narcissism, whereas supervision, warning and presence of administrator where listed as important situational aspects. Without a shadow of doubt the list given by Ziegler (2011) is not complete and it cannot be completed without further studies. In Figure 3 below there are some propositions of such characteristics, however, they should be treated only as such (propositions) and no firm conclusions should be drawn from the list presented. It is also worthy to remember that collaterally to Ziegler, similar model of factors influencing response biases was conceptualised by Leite and Cooper (2010) who formulated SDR as a result of a three-way interaction of person, situation and item characteristics. Similar ideas were already signalised by Furnham (1986) and Kalton and Schuman (1982).

Ziegler (2011) also urged for studies on process models of response biases. These studies are still lacking, that not only impedes verification of theoretical models, including satisficing and cognitive stages of responding, but also hampers research on methods to prevent and control for response biases. Ziegler (2011) compared it to developing treatments of a virus without knowing its exact nature. The field especially abounds with "faking" studies that are aimed at preventing and controlling for positivity bias in high-stakes contexts and are mainly concentrated on personality scales. However, it is not clear how faking studies inform item answering "under normal conditions" as context and content can influence not only the effect (bias), but also the processes leading to it (cf. Bensch, 2018).

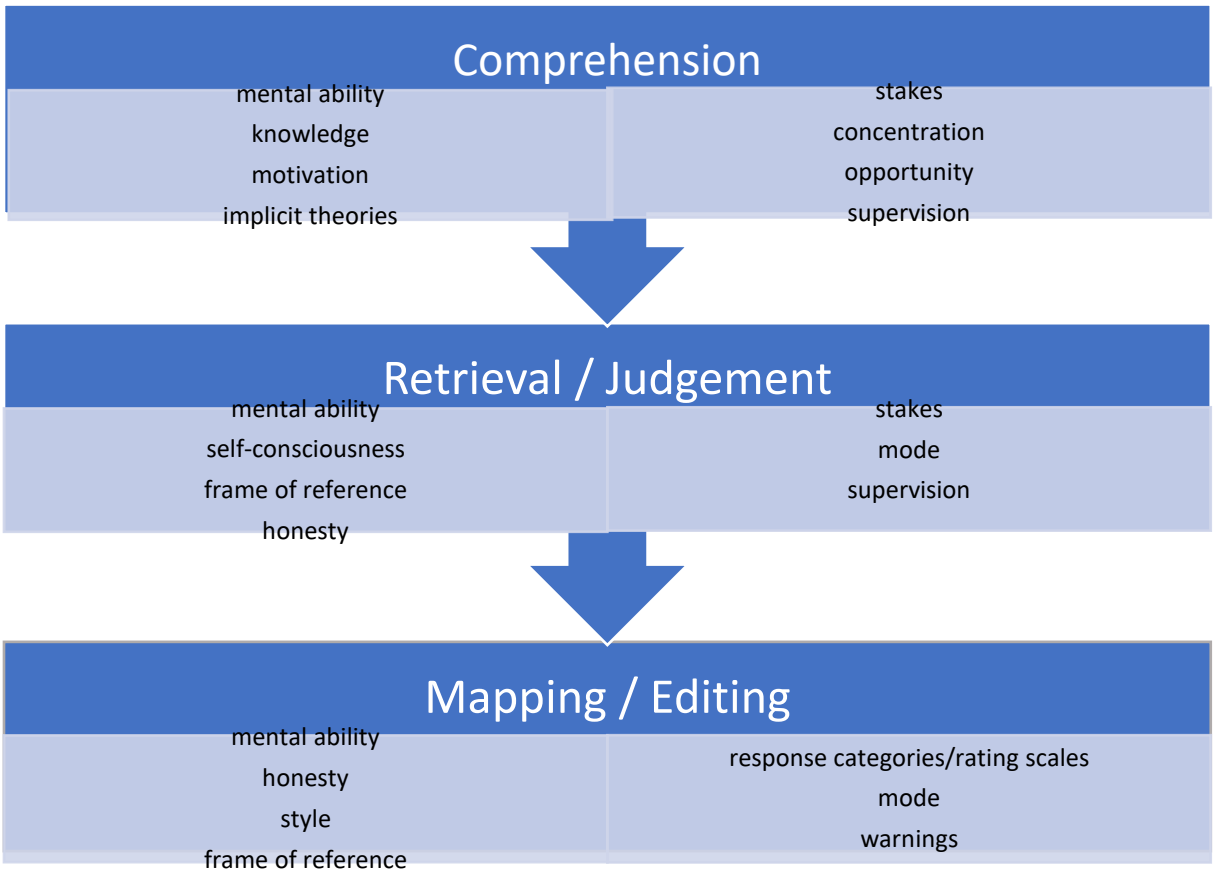

Figure 3. Cognitive stages of survey responding with person and situation characteristics influencing them (on the basis of Krosnick, 1999; Tourangeau & Rasinski, 1988; Tourangeau, Rips & Rasinski, 2000; Ziegler, 2011). Person characteristics are on the left, situation features- on the right.

This means that there is a great need of studies in low-stakes contexts as this is the most common setting for basic research measurements and also the only context in large-scale assessment programmes in the type of PISA, PIAAC, etc. These research efforts are especially needed as theoretical interpretation and practical utility of the most commonly used methods to measure positivity bias, e.g. SDR scales, has been recently questioned (Bensch et al., 2017). These scales, e.g. the BIDR, were shown to have a huge overlap with personality traits, hence, a question arises what is measured by them (Smith, 1997)? Is it a tendency to bias or a personality trait on its own (Bensch, 2018; see also Uziel, 2010)?

The forthcoming chapters will discuss various conceptions that can shed light on processes leading to positivity bias. The review is aimed to link SDR research and investigations related to self-presentation, self-enhancement, self-esteem, self-knowledge, self-consciousness and overclaiming. This review is also thought as a source of inspirations for new research questions and alternative explanations of positivity bias. Most of the theories reviewed have not been linked with response biases before or was linked only superficially.

Such a review is much needed as insufficient integration of research results within and across disciplines is a major problem for the response bias research. Even inside one subdiscipline (e.g. social psychology) there are a lot of similar constructs that are not sufficiently defined and differentiated from their synonymic counterparts, the same is true for cross-discipline research. Psychology, sociology and psychometrics can inspire each other and work together to solve response biases matter, but to this end greater synthesis between them is needed. Such a synthesis would help to join loose ends in current theories and verify some long-accepted truths that do not seem to hold in the face of new arguments. Uziel's discussion against the commonly accepted views on SDR scales is one of the examples (2010). Integration of approaches seems necessary to get a full grasp on mechanisms of response biases. Without that the efforts to find efficient methods to control response biases will be no more than just haphazard attempts to understand the outcomes without any insight into the "black box" (Ziegler, 2011). A comprehensive review of SDR control and prevention methods will be also presented in order to show why the overclaiming method is linked with such high hopes of response bias researchers.

2.3 Chapter summary

Self-knowledge and false judgments about the self were important philosophical topics already beginning from the antiquity, where they were present in Greek or Chinese philosophical writings. Interestingly, already these early thinkers considered self-knowledge to be the most difficult kind of knowledge attainable. Moreover, also the research on self-cognition biases and fallacies can be dated back at least to these remote times. More modern early researchers, like Charles Cooley or William James, complemented these thoughts by pointing to the role of social relations in forming and formulating self-judgments.

Early research attempts also developed a set of methods that was used in order to construct knowledge about the self, which included introspection, self-inference, self-verification, looking glass (inferring on the self on the basis of opinions held by others on our topic), social comparisons and self-perception, defined as culling information about the self on the basis of behaviour. Along with the methods' toolbox, a set of errors, biases, fallacies and illusions was identified, using the names given to these phenomena later on, both self-deception and faking/impression management were identified early on in the research history.

However, initial research attempts were characterised by a significant bend towards practical aspects (e.g. better measurement), somewhat ignoring theoretical development of the self-reports biases. The

concept of responding according to socially held norms was formed early on, but at first socially desirable responding (SDR) was seen as close to pathology, disturbance or delinquency. This psychopathological conception of the early SDR research was somewhat balanced by the Cronbach's ideas of response sets and response styles as main response biases. The former were thought of as transient, context specific and non-generalisable measurement distortions, whereas the latter as stable, generalisable and context-independent ways of responding that were linked with trait-like explanations. This trait-like conceptualisation of SDR prevailed in the later period of the history of this research topic and resulted in creation of numerous new methods to measure SDR: self-report scales (inventories) aimed at measuring propensity to respond according to social norms as an inter-individually varied personality-like trait. One of the first of such scales was proposed by Edwards, but shortly his questionnaire was criticised by Marlowe and Crowne who constructed (in 1960) the first SDR scale conceptualising this construct outside of the clinical framework. Their scale became a principal measurement tool to assess SDR for the next 40 years of research.

The field was greatly agitated by the watershed integration prepared by Damarin and Messick. Their work indicated that SDR comes in many aspects and that there are many relevant stages between conscious accuracy and conscious misinterpretation (inaccuracy, lying, faking). This meant that SDR was firmly defined as a heterogenous construct, moreover, the two researchers pointed that "bias in self-regard" (self-deception) posed greater measurement threat than "bias in self-report" (impression management) which was a big break-through in the field. Damarin and Messick for the first time suggested that SDR is a validity menace also in low-stakes, non-threatening and anonymous measurement contexts. This line of research was continued by Sackeim and Gur who asked why people were even capable of self-deception. They have concluded that self-deception was fuelled by motivational, which does not mean intentional, forces. Later on, Paulhus turned attention into adaptive advantages of self-deception, which can serve as self-esteem protection and facilitator, at least for short term, of social relations. Paulhus also proposed new classification of SDR types with impression management defined as a response set, with no trait variance and being entirely context-dependent. He also merged the framework of agency and communion (two super-motives organising social cognition and behaviour) with the SDR field by pointing that agency and communion organise content of self-presentation in a wide spectrum of social situations. Paulhus also corroborated earlier findings that the scope of self-deception is much wider than sexual, aggressive and in general socially inappropriate content. He also contended that impression management is a habitual, autotelic social behaviour, hence it is not necessarily defined as conscious or intentional lying and faking. Paulhus was also responsible for presenting a long list of SDR covariates, e.g. related to personality but also pointing to subjective importance of domain measured as an important predictor of SDR tendencies.

Later conceptions, e.g. formulated by McFarland and Ryan and by Ziegler, indicated that all SDR-related biases should be merged under one framework of the overly-positive bias of self-perception (positivity bias). New models of positivity bias (and other response biases) claimed that they are driven by a three-way interaction between person, situation and item characteristics of a given measurement occasion. Newer research again called for more response bias studies in low-stakes contexts. Biases in these situations are rarely studied but it is firmly evidenced that they are not absolved from response biases.

Chapter 3- POSITIVITY BIAS: RELATED CONCEPTS AND ALTERNATIVE EXPLANATIONS

3.1 Impression management and self-presentation

Impression managing and regulating is most often classified into the self-presentation framework and defined as tailoring one's impression to social reality and situational context (Schlenker & Weigold, 1992). Most often it is achieved by tactical selection of information- positive traits are overstated, negative ones are understated (Goffman, 1959). The problem of self-presentation attracted interest of social sciences (psychology, sociology, anthropology) already in the beginnings of the 20th century, when Charles Cooley wrote that mental images of self and others are social facts that influence inter-personal interactions (1902). Moreover, he also introduced to the James' theory of the self the ability to reflect over one's behaviour. Cooley also thought of self-image as being shaped predominantly by the perceptions of how others perceive an individual. In this conception, named looking-glass self, the self-perception was in fact shaped by social judgments based on self-presentational image that an individual presented to others. Thus, self-perception is to a large degree moulded in social interactions. At the beginning self-presentational behaviour did not have particularly high opinion and was rather considered as lying or a symptom of clinical or adjustment problems. However, the later research showed that self-presentation is in fact an adaptational behaviour that, to different degrees, is evinced by everyone (Leary, 1995).

3.1.1 Erving Goffman's views on self-presentation

Canadian-born sociologist Erving Goffman is believed to be one of the first researchers that came out with ground-breaking views on self-presentation. In his seminal book *The Presentation of Self in Everyday Life* (1959) Goffman stated that the social interactions are driven by "veneer of consensus", "surface of agreement", everything to achieve the "interactional *modus vivendi*" (p. 9). Goffman used a theatre metaphor to describe social situations when individuals ("actors") present their images for observers ("audience") by playing their "parts" in given "instants" (p. 16). In his conception each participant (actor) of an interaction participates in coining a "real agreement" as to "whose claims concerning what issues will be temporarily honoured" (p. 10), this leads to elaboration of a "working consensus", which has different content in each situation or setting, but which general form remains similar (p. 10). Goffman emphasised the importance of initial information ("first impressions") as they define the situation and their subsequent developments. Many of his interlocutors called this "getting off on the right foot", as in the example of a teacher, that advised to be strict for pupils at the beginning of work with them in order to build authority (p. 11-12). Moreover, the "conception of oneself" is an important initial factor defining the social situation. Self-presentation is thus a key to build this social consensus for a given situation, hence, when presentation is discredited, all participants feel uneasy (p. 12) and such disruptions can have serious social and individual consequences. Goffman even wrote about "social coin with awe on one side and shame on the other" (p. 70)- when "mystification" is disclosed, the performed may felt shamed[14].

In order to prevent such embarrassing situations, "defensive practices" are employed to "safeguard the impression" (p. 13-14). These "protective practices" are often called "tact" in the natural language

[14] Goffman claimed that observers have an inherent advantage over actors and that they check the informant, e.g. observe different aspects of his/her behaviour by checking up "on the more controllable aspects of behaviour by means of the less controllable" (p.7).

and prevent the definition of the situation from being abruptly changed. The Canadian sociologist also changed the views on self-presentation by pointing out that the clear-cut, blatant lies are rare, because they would be easy to disclose and would thus threaten the whole defined situation. Instead, omissions, intentional ambiguity, innuendo, over- and understatement are encountered much more often (p. 62). He also differentiated two kinds of communication in self-presentation: deceit and feigning. The former one is everything related to the communication *sensu stricto*, e.g. verbal utterances or gestures, the latter one is defined as performing actions that have other aim besides being informative, e.g. a worker can present himself as diligent to his supervisor by simply doing some job-related actions. Goffman called the whole process "misrepresentation"- one of the key elements of self-presentation in his conception. Therefore, these conceptions transformed self-presentation from a marginal process close to pathology into a general phenomenon, a cornerstone of social interactions. Goffman coined the term "impression management" and claimed that it happens everywhere in every social interaction (p. 15). He even claimed that this "game" is indispensable for the proper society's functioning (1959).

Obviously, impression management is not consisted of only true elements. The "mask" (self-presentation) presented to the "audience" is a "truer self", "the self we would like to be" (p. 19). Creating this impression of "better self" is a socially desirable presentation, called "idealisation" by Goffman, another key element of self-presentational behaviour. Cooley saw similar processes as a training for being better (p.35)- one tries to live up to one's ideals, presents orientation upward, celebrates norms of the society (1922). Negative idealisation is also possible, as an example Goffman gives a "poverty show" presented by street beggars in order to collect bigger alms (p.40). Idealisation can employ material or non-material values, e.g. showing off one's new car or performing certain behaviours that characterise only privileged members of society, e.g. higher Indian castes (e.g. Indian Brahmins). Idealisation can be based on showing something or concealing something, e.g. Brahmins perform self-presentation by drinking and eating in a specific to their cast way but simultaneously they may conceal fact of drinking alcohol or eating meat as these behaviours are not allowed for the members of this caste. Hence, Goffman differentiated very similar types of socially desirable presentation as the main strand of this research, namely enhancement and denial. He also pointed to strategic and dark secrets as things that individuals are especially unlikely to disclose. This part of Goffman's research refers to similar behaviours as the sensitive questions survey research.

Sometimes the participants grow to the presented image to such an extent that they internalise the role, e.g. a recruit in the army may be at the beginning bothered by military drill but with the time he accepts the rules and starts to live by them as played roles influence one's personality and become a second nature. Hence, images and masks are not things and faces, but they convey truth about them. Goffman claimed that IM can only be successful if it is "in sync" with individual's perception of self and that successful self-presentation leads to an image of "a self", this self being a product of the presentation, not its cause (p. 252). He also discussed the very much debated issue of conscious or unconscious character of IM. Goffman believed that is almost always both as an individual defines the situation "knowingly and unwittingly" and IM is hence a melange of controlled and automatic processes as "we all act better than we know how" (p. 74). Moreover, it is a "mixture of cynicism and belief"- self-presentation is partly true, but there is also, apart from self-presentational behaviour addressed to others ("the audience"), (self-)deception involved, as in the case of shamans that are convinced about their skills, though they also use "sleight-of-hand" tricks (p. 21). Hence, Goffman perceives "belief" in the presentation as another key element of successful IM. Presentation may be

sincere or cynical and can have agentic or communal motives[15] (p. 17-18) and can be triggered by an individual motivation or by traditions and expectations regarding one's group role, e.g. social status, group membership. He believed that individual's personality, given social interaction and given society all influence the precise motives of IM.

In case of each act of IM both "front", defined as physical appearance, decoration and "dramatic realisation", actions undertaken to give certain impression are used, but the information outflow is tuned to the social situation and individuals engaged, e.g. the presence of others changes the presentation like workers normally do not use familiar names or jokes when clients are present. Appearance of an "intruders" also change the presentation, e.g. sudden appearance of a supervisor may spur the workers brigade to drop leisure activities and get back to work. Goffman also pointed out that self-presentation is often performed in teams, e.g. all workers in an office may be engaged in a performance before clients or supervisors.

Goffman independently reached very similar conclusions about impression management and self-presentation as the SDR research conducted before and after his first book publication (1959). He claimed that self-presentation is an important and general social phenomenon, indispensable and inevitable in almost every social interaction. He also noticed that the character of IM can be both conscious and unconscious, and that probably most often it is a mixture of carefully planned and totally unwitting processes. Moreover, he sees IM as an entirely motivated process, based on motives closely resembling the agency-communion framework. Creating and maintaining positive (self-)image ("face") is realised through overstating positive characteristics (enhancement) and downplaying the negative ones (denial). Goffman also noticed that self-presentation can change quickly and without apparent reason, he ascribed it to an "everchanging character of the human nature" (p. 56) that can influence the content and degree of IM in a seemingly unchanged social setting. However, it has to be remembered that Goffman studied a related, but also very distant phenomenon from bias in questionnaire items. As he analysed behaviours that were performed in the immediate, physical presence of all participants (p. 15) he was forced to use qualitative methodology (observation), that is rather far from the quantitative character of the data that is in the centre of this dissertation. Nonetheless, having this caveat in mind, Goffman's work is an important contribution to the main frame of the SDR research and can be treated as a corroboration of many research ideas. Also, last but not least, Goffman coined the term "impression management" that was later adapted to the SDR field by Delroy Paulhus.

3.1.2 Impression management in Peter Blau's sociology

The concepts of social approval and impression management are also present in the research of Peter M. Blau, who reviewed their roles in social exchanges (1964). Blau considered social approval as one of the main motors of social behaviours as it was called "basic reward" in social interactions (p. 17). Moreover, winning social approval was seen by Blau as a key condition to find acceptance in any social group (p. 34). Therefore, people have to "prove attractive associates" in order to secure favourable position in the social network (p. 34). Despite the commonness and importance of this process of impressing others (impression management) Blau stated that "people create impressions continually and without special design", but also that a specific motive may push them to create impressions in a self-conscious and deliberate way (p. 39). It was also emphasised that the process of winning social approval is highly varied among individuals, groups and situations (p. 39). Interestingly, Blau noted that people often use a double strategy of being impressive and self-depreciating to win social approval (p.

[15] Obviously, as writing in late 1950s., Goffman did not use these terms precisely. He instead wrote about "private gain" and "the good of the community".

57). Employing self-depreciating tactics is especially important to maintain group integration and hierarchy. Probably self-enhancement has to be balanced by self-depreciation as social relations are vulnerable to vaunt and bloat. Superiors and subordinates yield different patterns of self-enhancement and self-depreciation- superiors are more modest in case of unimportant characteristics but remain high self-ratings in traits important for their position and group's success, whereas subordinates tend to self-depreciate their important characteristics in order to keep low, unthreatening profile towards their superiors. However, in order to compensate for this humbleness subordinates turn to self-enhancement in case of unimportant attributes. Subordinates also tend to conform with the views presented by their superiors (Blau, 1964, p. 54-56; Jones, Gergen & Jones, 1963). Blau also noticed that it suffices for approval of abilities to be unilateral (respect), whereas approval of opinions or beliefs has to be reciprocal (agreement). This leads to the forming of the social consensus that also helps to sustain one's view of himself. In this way social approval also shapes self-knowledge and influences changes in the public opinion (Blau, 1964, p. 62; Noelle-Neumann, 1974). According to Blau's observations, people are more prone to express approval than disapproval (p. 68), which may play its role in the commonness of self-enhancement tendencies- as mainly positive feedback is received it may lead to formation of overly positive self-views. However, only sincere social approval has value and the simulated one is not only worthless, but can also draw negative social consequences on the individuals using ineffective ingratiation tactics (Blau, 1964, p. 82).

3.1.3 Social approval and surveys on public opinions

Elisabeth Noelle-Neumann picked up one of the most interesting threads in Blau's research- how social approval shapes public opinions and leads to the views homogenisation on the group level and to the internalisation of the socially approved views on the individual level. Her research concentrates around the view that social approval leads to the forming of social consensus and is one of the basic forces shaping group processes (1974). People tend to conform to the most prominent or most common views in order to prevent from social isolation and sanction. Noelle-Neumann referred to the research of Asch (1951), Goffman (1959) and Milgram (1961) on conformism in order to conclude that people value social integration higher than their own judgements and views (1974). In order to gauge what is the prevailing opinion or norm they use a "quasi-statistical organ"- they observe their social environment and try to assess the prevalence of opinions and ideas and also evaluate the chances of success (chances to become uniformly accepted) of given norms or viewpoints (Noelle-Neumann, 1974; 1991). These processes lead to the steady homogenisation of the public opinions on a certain topic, in the aftermath there is only one commonly accepted and socially approved viewpoint on a certain matter. Those that do not agree with the dominating opinion are forced to remain silent or to fake agreement under the threat of social sanctions (e.g. isolation). In this way, through conformism, socially desirable responses in surveys are given. It is worthy to note, that this conception is the closest to the Crowne and Marlowe's need of social approval idea (1960), but puts much less emphasis on clinical and psychological needs and more on social processes. The process of forming social consensus was called "the spiral of silence"- as the viewpoint becomes less and less popular it is also less and less accepted in the society (Noelle-Neumann, 1974). People are swayed to conformity not only by social sanctions but also by the wish to protect self-confidence and self-esteem- people just want to be sure that they are "right" in important issues. Noelle-Neumann also identified correlates of opinion conformity- women, elderly and people of lower socio-educational status were more prone to conform, whereas men, younger and those of higher socio-economic status were more prone to speak up and oppose the common views (1974). It is striking that the same groups, especially older and less educated participants, are also often linked with lower quality responses in self-report research (Grau et al., 2019; Magdolen, von Behren, Hobusch, Chlond & Vortisch, 2019). Hence, it can be assumed that conformity, the spiral of silence and SDR are driven by very similar mechanisms. However, the research

also shows that almost always there are some groups that resist the common opinion and stick to their own views even when faced with social isolation or odium. It is unclear what are the correlates and mechanisms of forming such "hard-core" partisans of lost cases (Noelle-Neumann, 1991). Noelle-Neumann noticed that selecting informational sources, nowadays called most often "information bubbles" or "filter bubbles", are most probable reasons of this recalcitrance (1974)[16].

Interestingly, the "quasi-statistical organ" of measuring public opinion is often wrong as it is susceptible to biases (Noelle-Neumann, 1991). People fall prey to false-consensus bias- they tend to wrongly assume that their own views are more widely shared in the society than it is in reality. Moreover, social sanctions, e.g. isolation, are not the only negative consequences feared by conforming responders. They are also afraid of embarrassment and social derisiveness as potential consequences of yielding unacceptable or stupid opinions. Noelle-Neumann's research also corroborated the view that people form opinions about their own traits also one the basis of what others think about them (cf. Cooley, 1902; Taylor, 1982), but noted that these assessments also often tend to be wrong (1991).

Noelle-Neumann's research on SDR and yielding conforming answers was further developed by Bishop and his co-workers in their research on pseudo-opinions (1986). These researchers proved that people can not only refrain from giving responses if they feel unsure of whether they "should" think on a given topic. Apparently, they are also prone to give opinions on fictitious, non-existing issues if they think they should do so (Bishop, Oldendick, Tuchfarber & Bennett, 1980; Bishop, Tuchfarber & Oldendick, 1986). Bishop and his colleagues offered a survey where participants were asked to provide their opinions on various current political issues. Among the real items there were few non-existing issues, e.g. "Metallic Metals Act" (Bishop et al., 1980). To their surprise, about one-third of the sample offered their agreement or disagreement with the fictitious items. Moreover, in a split-ballot design some participants still persisted on giving an opinion on non-existing items even if offered an explicit occasion to say they had no opinion whatsoever (filtering question; Bishop et al., 1986). Participants that were men, less educated and claiming less knowledge about politics in general were more prone to offer opinions on fictitious issues, however, this effect disappeared in filtered condition (Bishop et al., 1980; 1986). The effect was interpreted in a way that less educated participants yield opinions on non-issues because they want to pose as more knowledgeable than there are in reality. It is possible that this behaviour is a self-esteem saving mechanism as by yielding pseudo-opinions participants are claiming "I am not uninformed. I have an opinion!". In the filtered condition however, they may feel relieved from this burden, as the option to provide an "I don't know" opinion is explicitly provided to them. In this condition a pressure to offer such opinions could be greater among better educated participants as they feel obliged to be well-oriented in politics due to their superior status. It is interesting however, that Bishop and his colleagues did not find any relation between propensity to offer non-opinions and score on the MCSDS (1980; 1986). Hence, survey participants may not only claim socially desirable characteristics or refrain from yielding responses putting them in an unfavourable light, but they may also offer opinions on non-existing, completely fictitious items if they only think they should do it in order to yield desirable or expected impression. Nevertheless, lack of correlation between propensity to yield such pseudo-opinions and the MCSDS challenges simple ascription of pseudo-opinions to SDR processes[17].

[16] See also e.g. Cornelis, Van Hiel, Roets & Kossowska (2009) for personality and socio-demographic correlates of conservatism and Kappes, Harvey, Lohrenz, Montague & Sharot (2020) for socio-cognitive and neuronal basis of confirmation bias and altering opinions. This part of the Noelle-Neumann's work is however beyond the scope of this dissertation so the topic will not be continued here.

[17] Or challenges the MCSDS construct validity as a measure of SDR (Smith, 1997).

3.1.4 Robert Hogan's socioanalytic theory

An interesting consequence of the integration of the SDR research with other research fields from sociology and social psychology is a steady departure from the inconvenient terms of conscious and unconscious bias in favour of automatic versus controlled self-presentation. This change goes in line with the Hogan's "socioanalytic" theory (1983) according to which constant practice of public self-presentation leads to automatization of the process and results in habitual self-presentations even in private contexts. According to the Hogan's ideas, a controlled self-presentation directed towards a given audience equals frank self-descriptions performed in private settings. In this way the self-reports in "private audience" settings may still be biased due to self-deception and habitualised self-presentation. Adapting this line of thinking enables to support claims of a domain- and context-general character of self-deception proposed by Paulhus (2002) who expanded the scope of this action out of the purely self-defensive character. Traditionally self-deception was thought of as a motivated action, aimed at e.g. self-esteem protection, but according to the Paulhus' interpretation of the Hogan's ideas, self-deception may also have a non-motivational, automatic character (2002).

Hogan based his theory on social psychology, evolutionary biology and symbolic interactionism, being heavily influenced by Goffman's work. However, unlike Goffman, Hogan sees social roles and self-presentational behaviours related to them as products of inner self-concept, not outer social forces. In other words, according to Hogan and on the contrary to Goffman or Gergen and Gergen (1980), it is the self-concept that choses the roles it wants to play in social interactions, not the other way around (that the roles played shape the self-concept). Hogan believes that self-concept is organised around the core that is formed in social interactions from infancy to adulthood and that also has a significant biological component (Hogan, Jones & Cheek, 1985). The self-concept can change in the course of life, but Hogan opposes views that self-concept is constructed *ad hoc*, according to the social role actually played. Hence, Hogan's conception entails that inter-situationally changing self-presentations are always based on the stable self-concept. Hogan also implies that all social actions are driven by three "master motives": a) need of approval, approbation and attention (need to get along, clearly related to communality motives), b) need of power, status (need to get ahead, obviously related to agentic motives) and c) need of structure, order, predictability. According to Hogan "master motives cause and explain social actions" (Hogan et al., 1985). All motives are important for an effective social group functioning, especially the balance between getting ahead and getting along must be observed in order to preserve group cohesion. The socioanalytic theory states that people are highly inter-individually varied in case of how they pursue these three master motives. Most importantly to the central topics of this work people vary in case of: a) what they perceive as socially desirable, b) what behaviours they yield in order to appear socially desirable, c) importance they ascribe to respond in socially desirable way (Johnson & Hogan, 2006). The first difference determines content and context of self-presentation, the second one defines the exact behaviours used to achieve goals and the third one stipulates the degree of self-presentational actions, in some cases agents may decide not to engage in such behaviours at all. From the point of view of this theory any testing situation, be it answering personality items or participating in an online survey, is just another social situation where exactly the same rules apply. Participants SDR is influenced by their motivation to provide socially desirable images and also their skill to do so (Johnson & Hogan, 2006). Most often people yield responses that are a resultant between their identity ("who am I") and their reputation ("how others see me"). Hogan states that the responses given may be valid (not distorted in socially desirable way) even if they do not closely resemble the actual state of affairs in the real life. It suffices if they convey the social image of a participant validly (Johnson & Hogan, 2006). This statement is based on the concept that both identity and reputation convey certain "truth" about one's characteristics: "reputation describes a person's behaviour; identity explains it" (Hogan & Holland, 2003).

3.1.5 Other views on impression management

Therefore, impression management was established as an omnipresent phenomenon concerning every sphere of life, including eating, entertainment or attire, being so important for certain people that to maintain a given presentation they are ready to risk their life or health (Leary, Tchividjian & Kraxberger, 1994). Further research concentrated on the functions of IM. At first, it was commonly accepted that the main function was gaining social acceptance[18] (Schlenker, 1980), but afterwards this view was revised and three main functions of IM were determined as: a) interpersonal influence (e.g. eliciting a certain impression to achieve an aim), b) building own self-identity and self-esteem and c) emotion regulation (Leary, 1995). It is noteworthy, that IM can serve to achieve both social as well as individual goals, e.g. protecting self-esteem, maintaining identity (Gollwitzer, 1986).

Impression management was proven to depend on three groups of factors: a) social norms, roles and values, b) already possessed social image and c) a precise motivation (goal to be achieved) by the dint of IM (Bond, 1991; Leary, 1995). Individual motivations of IM are inter-individually varied to a great degree and similar variability characterises competences to convey a desired impression. High motivation to manage social impressions is related to character of interaction (public vs. private), importance of the desired goal and difference between the desired image and the image perceived by observers (the greater the discrepancy, the greater the IM). Motivation to IM is also elevated when future social interactions with a given audience are likely in future.

Self-consciousness and self-monitoring

Tendency to IM, treated as a trait, is also individually varied and two trait-like constructs were proposed to describe it. The first one is public self-consciousness- high levels of this trait mean paying larger attention to the public, social aspects of one's image and are also related with higher levels of conformity as well as large anxiety of social rejection (Fenigstein, Scheier & Buss, 1975). The second one is self-monitoring and it is understood as an ability to control and manage own social impressions and to tune them to social situations basing on the guidance from social observations and feedback given by the audience. People high in self-monitoring can precisely identify and interpret even subtle signals from social interactions and consequently tune one's IM to them (Snyder, 1974). Self-monitoring ability is also linked with efficient self-control of emotion regulation and able communication through various channels (facial, vocal, etc.). Snyder proposed that self-presentation can be driven either by self-monitoring (social observation, self-control, attunement to social appropriateness) or by inner emotional states. Efficient self-monitoring involves also paying great attention to social comparisons which are treated as a primary source of information (guidelines) on how to prepare self-presentation for a given social situation. Thus, to effectively self-monitor means to control one's own behaviour and emotional expression and to observe other people in the search for social information needed to manage the right impression for a given situation. The ability to self-monitor is believed to be highly inter-individually varied and that "monitored expressive behaviour should vary more from situation to situation than non-monitored expressive behaviour". Thus, in case of effective self-monitors their behaviour is more adjusted to the social situation whereas in case of low self-monitors their actions are just driven by their emotional states and stable personality traits. Snyder presented empirical results that showed that professional actors have higher self-monitoring abilities than university students and that university students have supreme self-monitoring skills than psychiatric patients. Self-monitoring is considered a learnt ability, e.g. in the process of socialisation, that it cannot be just simply inherited (1974). Self-monitoring was also conceptually contrasted with SDR: the former is a skill, an ability, and the latter, in the Snyder's view of SDR based on the approach

[18] Cf. Marlowe and Crown's theory of need of social approval as a main drive of social desirability (1960).

of Marlowe and Crown, is a motivation to seek for social approval. In later research Arkin (1981) and Rauthmann (2011) proposed that self-monitoring can be divided into acquisitive (offensive) and protective (defensive) self-monitoring, in a similar way as Leary (1995) divided self-presentation. Similarly, Barnes (1976) suggested that impression management can be divided into assimilative IM and accommodative IM, the former being bringing other people behaviours in line with one's own expectations, the latter being somewhat opposite- adjusting one's behaviours to the views of others (Gangestad & Snyder, 2000). Briggs and Cheek (1988) questioned self-monitoring theory, accusing it of being redundant to extraversion, but Gangestad and Snyder (2000) refuted this claim, by proving that the two concepts, although closely related, form separate traits, as extraversion does not contain sensitivity to others and interpersonal orientation to such a degree as self-monitoring. It is to be determined by further research, whether self-monitoring is just a facet of extraversion, or is it a standalone personality trait.

Related concepts: Dark Triad personality and overconfidence

Another trait similar to self-monitoring is Machiavellianism, defined as tendency to conscious and strategic manipulation of interpersonal relations, including deliberate lie, deceit and distorting self-image (Christie & Geis, 1970). Machiavellianism is also connected with emotional coldness and indifference to morality (Jones & Paulhus, 2009). Other "Dark Triad" personality traits: narcissism and subclinical psychopathology also received research interest in the context of their relation to SDR (Paulhus & Williams, 2002). Narcissism is a personality trait related to SDR because of the proneness to exaggerate, boast and conceit characteristic for subclinical narcissists (Raskin & Hall, 1981). Subclinical psychopathology is related to low levels of anxiety and empathy along with high degrees of thrill-seeking and impulsivity (Hare, 1985).

However, as much as the above-mentioned constructs seem related to the SDR, they are nevertheless distinct from it, both conceptually and empirically (Paulhus, 2002). It was evidenced that self-monitoring did not correlate with SDR scales: Paulhus (1984) correlated Snyder's self-monitoring scale with popular SDR scales and did not receive significant correlations. Similar results were obtained by Snyder himself who did not obtain substantive correlations between his self-monitoring scale and the MCSDS (1974). More results on that were presented by Kowalski, Rogoza, Vernon & Schermer (2018) who correlated various measures of the above-mentioned constructs. In their research self-monitoring was not related to SDR but was moderately and positively correlated with all Dark Triads traits, noting the highest correlation with narcissism. SDR, on the other hand, noted weak correlations with narcissism (positive), Machiavellianism and psychopathy (both negative). All Dark Triad traits yielded moderate and positive correlations with each other (Kowalski et al., 2018). Because Kowalski and his colleagues did not provide any path analyses or structural models nor did they perform a possible suppression/mediation analyses it is impossible to tell why these constructs correlate with each other in the obtained fashion. Moreover, they have used a less-researched SDR measurement tool instead of the most commonly used the MCSDS and the BIDR. More research is clearly needed in this field. However, for the needs of this dissertation it suffices to conclude that SDR, as measured by SDR scales, is only very distantly related to self-monitoring and Dark Triad personality traits. The only clear relation that can be explained on the conceptual level is the positive relation between narcissism and self-enhancement tendencies, as both contain grandeur illusions (Paulhus & Williams, 2002).

Overconfidence is yet another topic that is often related to SDR. In some conceptions it is even placed in the common core of the positivity bias along with SDR and overclaiming (Ziegler, Maass, Griffith & Gammon, 2015). The most common definition of overconfidence is "a positive deviation between estimated and actual performance" (Stankov & Crawford, 1997). This concept can be further divided into: a) overestimation- thinking to be better than one really is, b) overplacement- thinking to be better

than others, c) overprecision- being more sure of one's knowledge that it is warranted (Moore & Healy, 2008). A recent meta-analysis shows that confidence is in general related to the correctness and speed of response, indicating that people are able to predict the results of their actions in mnemonic or perceptual cognitive tasks. Only around 6% of participants is characterised by a negative prediction that is similar to overconfidence (Rahnev et al., 2020). No stable correlates of overconfidence were identified as yet, apart from narcissism and extraversion being related to overprecision (Moore & Dev, 2017). Despite the superficial similarity to SDR there is no evidence for a larger relatedness of the two phenomena (Bensch et al., 2017). Any correlations between SDR and overconfidence identified to date never exceeded correlations around 0.30. It is thus believed that overconfidence is a part of different nomological network and that it is rather a metacognitive skill than a response bias (Kleitman & Stankov, 2007). It is possible that future research will help to build the nomological network of overconfidence as some of it measures are under heavy critique (Moore & Dev, 2017). However, overconfidence seems to be a topic that is at best only loosely related to SDR and will not be referred to further in this work.

3.2 Self-esteem and SDR in the survey research context

Self-esteem is another construct often linked with impression management and self-presentation. It is related not only to the degree of self-presentation, where lower self-esteem correlates with more self-presentational behaviours, but also to the content presented (type of impression conveyed). High self-esteem is associated with offensive (acquisitive) self-presentation where a lot of positive characteristics are ascribed to an individual. Offensive self-presentation is akin to self-enhancement and it is performed with a high belief in success (cf. Goffman, 1959). Low or unstable self-esteem is related to defensive (protective) self-presentation, where main motivation is to save oneself from negative presentation and to conceal negative traits (Leary, 1995). Low self-esteem is also linked with social anxiety and need for social approval (Leary, 1990). This construct is clearly distinct conceptually from SDR but from the early beginning of the SDR research the two concepts were intertwined in many research attempts, for example self-esteem protection was fundamental for many conceptions of SDR mechanisms (the so-called motivational accounts, e.g. Sackeim & Gur, 1978). Paulhus and Trapnell (2008) included this idea in their classification of SDR types based on the agency-communion framework: protecting self-esteem was to be based on agentic self-presentation, whereas avoiding social disapproval- on a communal one.

Another popular theory linking self-esteem and SDR contended that self-enhancement was beneficial for personal adjustment[19]. This view was formulated partially on the basis of correlations between SDR scales and self-esteem self-reports (Paulhus, 1998). Moreover, it is postulated that self-enhancement bias is actually one of the sources of self-esteem, as distorted views of oneself and one's accomplishments may lead to an elevated self-regard (Kwan, Kuang & Hui, 2009). Researchers also pointed that self-esteem may be an outcome of various biases: either narcissistic tendencies or self-enhancement, leading to different consequences for the level and stability of self-esteem. Narcissistic basis of self-esteem often results in maladjustment and negative, especially long-term, social consequences (Crocker & Park, 2003; Kernis, 2003; Paulhus, 1998; Tracy, Cheng, Robins & Trzesniewski, 2009). Self-enhancement, especially moderate in the degree of distortion does not seem to bring such negative consequences for self-esteem (Kwan et al., 2009). Other conceptions see self-enhancement as one of the main self-motives[20], powerful drives used to regulate and shape self-

[19] In this account, according to the established consensus, all the terms akin to self-esteem (e.g. self-worth, self-concept, self-regard, self-image, self-evaluation, etc. are treated as equivalent variants of the general term "self-esteem" (cf. Huang, 2013; Kozielecki, 1984; Szpitalak & Polczyk, 2015).

[20] Other self-motives are: self-verification, self-assessment and self-improvement (Sedikides, 1993).

esteem (Leary, 1999; 2003). Individuals with high self-esteem tend also to self-enhance, but according to some conceptions this motive characterises persons with high but also unstable self-esteem and in a long-term it may hinder their functioning by blocking the self-improvement motive (Kernis, 2003; Szpitalak & Polczyk, 2015).

Other important question concerns the relation between self-esteem stability and SDR. An unstable self-esteem relies on external sources of information to confirm itself (Kernis, 2003) and it is evidenced that individuals with high but unstable self-esteem yield socially desirable responses, whereas individuals with high and stable self-esteem report positive traits but without much concern for social desirability (Schneider & Turkat, 1975). However, in order to fully explore the relation between self-esteem and SDR much more research is needed, especially using measures that are not self-reports. This should enable to investigate other facets of self-esteem than just subjective individual's feelings and opinions which could provide additional information on behavioural manifestation of self-esteem's level and stability (Coopersmith, 1967).

Huang (2013) analysed relations between self-esteem and SDR in a detailed way. In a performed meta-analysis the previously suggested positive relations between self-esteem and SDR were investigated (e.g. Arlin, 1976; Astra & Singg, 2000; Lindeman & Verkasalo, 1995; Riketta, 2004) and the alleged moderating role of SDR for the criterion validity of self-esteem measures was tested (Zerbe & Paulhus, 1987). The results of the meta-analysis pointed out that self-esteem, mainly measured by the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965), and SDR, mainly measured by the MCSDS or the BIDR, correlated around 0.30, a correlation that can be classified as moderate at best according to the Cohen's classification of effect sizes (1988). Moreover, SDR scales proved to have only a minimum moderating effect on the criterion-related validity of self-esteem measures. These results confirm the positive relation between self-esteem and SDR, although the magnitude of the relation is much smaller than previously thought. Interestingly, self-esteem measures correlated more with the SDE subscale of the BIDR (0.40) than the IM subscale (0.16) of this scale. However, this result is most probably result of agentic content of both the RSES and the SDE scales (Gebauer, Sedikides, Verplanken & Mayo, 2012). It is also possible that these correlations were driven by more general self-enhancement tendencies that are known to have larger effect on agentic traits, at least in individualistic cultures (Kurman, 2001; but see e.g. Church et al., 2006; Yik, Bond & Paulhus, 1998 for contradicting results). SDR also failed to significantly moderate the relations of self-esteem measures with other important variables, e.g. academic achievement or job performance, thus partialling out the SDR variance (implied as spurious variance) did not lead to better estimates of these relations. Huang interprets these results in the light of the overly positive bias of perception theory, claiming that people have an inherent inclination towards assessing everything related to them more positively than other things (2013; see also Pedregon, Farley, Davis, Wood & Clark, 2012). Participants' age and gender failed to moderate the self-esteem-SDR relation (Huang, 2013), probably due to low correlations between participants' age and SDR (only around 0.10; Ones & Viswesvaran, 1998) and small relation between gender and SDR (around 0.20, with males yielding greater SDR scores than women; Ones & Viswesvaran, 1998).

3.3 Self-knowledge and self-consciousness

3.3.1 Main theories of self-knowledge and self-consciousness

Self-esteem is strictly related to other two popular concepts: self-knowledge and self-consciousness. Despite their popularity, e.g. in natural language, both these constructs are still deemed a largely under-researched area with a lot of research lacunas (Kozielecki, 1981; Zaborowski, 1989). Self-knowledge is defined as a set of judgements about self, including self-descriptions (e.g. "I am tall", "I live in the USA"), self-evaluations ("I am good at math"), self-standards ("How should I act? Who I want

to be in the future?"), rules about generating self-knowledge and rules guiding self-presentation, e.g. what kind of self-presentation is accepted by an individual (Kozielecki, 1981). These standards are believed to be internalised and subjective, some of them may be even not intersubjectively communicable. Self-knowledge is polyfunctional, serving cognitive, motivational and regulatory mechanisms. Self-knowledge enables identifying oneself ("Who am I"?) and knowing others by simulating others' minds ("What do they think?"), as well as controlling (self-control) and regulating (self-regulation) one's own actions ("What should I do now?", "I can't go for a party without doing my homework"). Self-knowledge is believed to be created in the process of autoperception (self-observation, Bem, 1972), social comparisons (Festinger, 1954) and internalising cultural norms (Markus & Kitayama, 1991). These conceptions where first voiced in the work of William James (1890) who saw sense-making and integrative processes as foundational for self-knowledge. He also formulated conception that self can be either subject or an object of cognition, using the Sanscrit terms of "atman" and "jivatman" to name an active subject reflexing over the world and itself ("I") and an object of this reflection ("Me"), respectively. These thoughts were further developed by G.H. Mead who claimed that the role-taking process that is continuously exercised in social interactions leads to development of self-conversations and an ability to simulate others' internal states (1934). These abilities then lead to formation of self-consciousness, the result of which, according to Mead, is creation of personality (self-knowledge) ascending as a product of continuous self-reflection and self-criticism. Hence, self-consciousness is a key to "know yourself" as it is an essential process that enables creation of self-knowledge. Interestingly, large individual differences are observed in the ability to be self-conscious, or rather self-aware of who and how we are (Fenigstein, Scheier & Buss, 1975; Zaborowski, 1989). Self-consciousness can be defined both as a skill and as a state. The former is an ability to achieve self-insight and self-assess one's behaviours or abilities accurately, the latter is a state of being aware of one's qualities, thinking about them knowingly (Fenigstein et al., 1975; Wicklund, 1975; Zaborowski, 1989). It is unclear what exactly can trigger autoreflection, under what conditions it emerges and when self-conscious "I" is activated (Zaborowski, 1989).

Some researchers limited self-consciousness to a state of autoconcentration that served to compare salient elements of the self with social or personal standards and to adjust or regulate the behaviour accordingly. In line of these theories self-consciousness is always unpleasant as it inevitably brings troublesome feelings stemming from the comparisons between the current "Me" and an ideal self (ideal version of "Me") or an ought self ("Me" I-should-be) (Duval & Wicklund, 1972; Higgins, 1987; Wicklund, 1975). These self-discrepancies were believed to be such unpleasant that they generalised over to any state of autoconcentration, even one's own reflection in the mirror or hearing one's voice on a tape recording (Sackeim & Gur, 1979). This approach has been heavily criticised as ignoring the processual character of self-consciousness and over-concentrating on negative emotions caused by the state of self-concentration (Zaborowski, 1989). Hull and Levy (1979) proposed an alternative theory of self-consciousness, accenting automatic and processual character of the concept, emphasising its role in coding information about the self that incomes from external sources. Scheier and Carver (1977), Buss (1980) as well as Csikszentmihalyi and Figurski (1982) were among the researchers who criticised the negativity brought by the self-conscious states and contended that these states simply strengthen the current emotions, positive and negative alike. Zaborowski (1989) formulated his conception of self-consciousness as an integrative account of the earlier theories by Wicklund (1975), Hull and Levy (1979) and Rogers (1961). First of all, Zaborowski's theory postulated a multifaceted, complex account of self-consciousness (S-C). In his opinion S-C can be characterised by different saliency levels, with the lowest called latent, difficult to concentrate on. S-C can be also differentiated regarding content, with both public (social) and private (personal) contents possible to be in the focus of attention, and form, with both objective and subjective processing attainable to concentrate on,

depending on the centrality of a given domain (cf. Buss, 1980; Hull & Levy, 1979). In its processual face S-C was responsible for coding, processing and integrating information about the self. Zaborowski also contended existence of different levels of scope of S-C, with the defensive level being the most limited, defined as concentrating on self-defence against threatening stimuli. This type of self-consciousness was attributed to the research of e.g. Wicklund (1975) that postulated that autoconcentration was aversive. Greenwald (1985) discerned three potentially threatening situations that engage the self ("ego" in his papers) and may lead to biased self-consciousness (self-insight): a) situation of being assessed by others, b) being assessed by the self (self/auto-evaluation, also known as self-assessment) and c) processing especially important or sensitive matters to the self.

Zaborowski, however, argued that self-consciousness was unpleasant only when dealing with threatening stimuli, in people with low self-esteem and in case of negative affective states experienced by an individual. He postulated that other, more developed and wider in scope, forms of S-C were: a) personal, embracing objects related to physical and psychic phenomena related to the self, b) external, consciousness of the self in social interactions, and, the widest in scope and the most developed form of S-C, c) meta self-consciousness alias reflexive self-consciousness. This last form of S-C was responsible for reflexing upon different aspects of self and steering one's behaviour and cognition, as well as correcting and developing them[21]. Thus, in the account of Zaborowski, self-consciousness serves to construct self-judgements, change self-esteem, apply self-control, alter social relations or manage motives, igniting ones and dampening others. Zaborowski postulated a modular conception of self-consciousness, seeing S-C mostly as a polyfunctional and multifaceted process leading to the formation of self-knowledge and having a decisive role in steering everyday behaviour and cognition (1989).

However, Zaborowski was also aware that self-consciousness only sometimes meant objective coding of incoming information about the self. More often, the process of acquiring self-knowledge is threatened by many biases. Greenwald (1980) went even that far as to claim that the self ("ego") is like a totalitarian regime that actively distorts and even fabricates self-knowledge by information-control processes. Other researchers, however, were more optimistic about the possibility of forming more veridical self-knowledge. George Kelly built his theory of personal constructs around the notion that people are like naïve scientists when constructing their knowledge about the self[22] (1955). Scientists, as they are motivated to have an accurate view of the reality and because they use scientific-like reasoning to build and organise their knowledge. Naïve, because the systems of constructs they create are often distorted by their "idiosyncratic experiences" and in result deviate from reality (Kelly, 1955). The research led by Rubinstein (1960) corroborated Kelly's views and showed that not in every case the contents of self-consciousness reflected the reality- in case of low educated participants the S-C was very limited, reduced, superficial and seemed to be created *ad hoc*, without much coherence, as if their self-insight was only interim. These remarks were then confirmed by Loevinger, Wessler and Redmore (1970) and by classical research of Alexandre Luria who in the mid-1920s interviewed Uzbekistani peasants to discover that they were unable to yield any coherent self-judgements nor to describe their "character" (1974; Kozielecki, 1981). Nisbett and Wilson (1977) and Kofta (1979) contended that people on average are self-conscious of the contents of their self-knowledge, but not of the processes and mechanisms leading to its creation, arguing that the rules of self-knowledge formation are difficult and obscured from the simple self-insight of most of the people.

[21] It is worthy to add that Łukaszewski (1974) also distinguished similar, meta-level, regulative functions of self-consciousness. This notion in the theory of Zaborowski is also partly inspired by the Rogers' (1961) conception of self-actualisation (getting better, developing one's potential, closing to one's ideal self).
[22] And the world in general.

Other research pointed out that self-knowledge in certain domains tends to be richer and more organised than in other domains (Gordon, 1968; Jones, Sensenig & Haley, 1974; Markus, 1977). In these central and important domains self-knowledge is arranged in coherent, hierarchical structures linked with faster, more precise processing of information and low tolerance for ambiguity or discrepancy. Hazel Markus (1977) called such organised chunks of self-knowledge self-schemata (self-schemas) and contended that they help individuals to predict and control their own actions, aid to process domain-related information, facilitate retrieving relevant evidence and enable maintaining coherent self-image in a given dimension. In other, less important domains, self-knowledge is believed to be less organised, consisted of isolated, atomised judgements, of low accessibility and verbalisation (Alschuler, Weinstein, Evans, Tamashiro & Smith, 1977; Kozielecki, 1981). Recent research showed that another important differentiation should be among context-dependent and context-independent self-schemata (Klein & Lax, 2010). They are functionally independent and have different neuro-correlates with context-dependent selves related to memory processes and context-independent to more abstract reasoning (Martial, Stawarczyk & D'Argembeau, 2018). Thus, it looks like self-knowledge consists of general self-representation ("Who am I?") and a set of context- or domain-specific "selves" that differ in the degree of development, e.g. a professional tennis player would be most probably characterised of a very well developed self-schemata in the domains of her tennis abilities and much less elaborated self-knowledge in less relevant domains, e.g. math skills.

The above-presented results show that self-consciousness is an indispensable process in order to inspect and communicate self-knowledge (achieve self-insight). The existence of metacognitive processes regulating acquiring and organising self-knowledge is also postulated. However, little is known how these regulative processes operate. Especially interesting, from the survey methodology point of view, would be to investigate the mechanisms of verifying and correcting self-insight veridity ("Am I really good at math?") before communicating self-relevant information, e.g. in a survey study. On this level of research advancement it is impossible to assess how similar or dissimilar are these processes in comparison to the metacognitive functions regulating more basic, perception-level cognitive processes. Judging about the perception processes, e.g. awareness and confidence of perceiving an object (Siedlecka et al., 2019a; Siedlecka et al., 2019b), is quite well studied as is error monitoring and adjustment in performing very basic actions (e.g. Botvinick, Cohen & Carter, 2004; Fazekas & Overgaard, 2018; Ridderinkhof, Ullsperger, Crone & Nieuwenhuis, 2004; Veen & Carter, 2006). How and if this research can be merged with the research on self-knowledge and self-consciousness of much more complex information-processing is yet to be determined.

3.3.2 Generating self-knowledge and self-judgements (self-descriptions)

However, these and similar results did not bring answers why a stable, consistent and easy-to-communicate self-knowledge seems to be a rare thing (Gordon & Gergen, 1968). In most of the cases, even in the case of central, important dimensions, self-knowledge is distorted by biases that lead to a creation of a self-knowledge system discrepant from reality. Some researchers (Mor & Winquist, 2002; Sackeim & Gur, 1978, 1979; Wicklund, 1975), ascribed this effect to a lack of motivation for a realistic and valid self-insight due to the alleged unpleasantness of self-concentration. Turner (1975) considered two alternative explanations: a) lack of motivation or b) lack of abilities to perform a valid self-insight. Simply put, adequate self-knowledge may not be of primary importance to some individuals and reaching exact self-knowledge is always related to a substantial cognitive effort. Probably, most of the people are just satisfied with only an approximate self-knowledge in most of the situations (Kozielecki, 1981). Nevertheless, subsequent research showed that processes of self-knowledge creation and communication are mainly based on three methods (strategies) of constructing judgements and that they are driven by four basic self-motives (Kozielecki, 1981;

Sedikides, 1993). Kozielecki distinguished linear strategy of constructing self-judgements that was based on processing every available piece of information on a given topic before forming a judgement. Another strategy was called "conjunctive"- an individual is comparing information and forms a negative self-judgement if he finds even one negative instance. A seemingly opposite strategy was called "alternative" where even finding one positive chunk of information suffices to form a positive self-judgement (1981; cf. Lewicka, 1978). Thus, if an individual would be asked to assess his math abilities, then using the linear method he would compare every available instance of a situation when his math abilities were tested and then weigh them to see how much evidence is for good and how much for bad math skills. On the other hand, using the conjunctive method would lead to browsing through instances informative about his math abilities and forming a negative judgement ("My math skills aren't very good") when even one chunk of negative evidence could be found (e.g. exam failure). In this sense the conjunctive method is a kind of "all or nothing" strategy. Finally, when using the alternative method an individual also reviews the evidence on math skills but if only a one bit suggesting positive evidence is found (e.g. exam success), a positive judgement about one's math skills is formulated. Thus, the alternative method is a "one suffices" strategy that can lead to overly optimistic judgements (overclaims), whereas the conjunctive strategy can contribute to overly pessimistic self-perceptions (underclaims).

Generating self-judgements is thought to be driven by four self-motives (auto-motives), differed in their function and strength (Bargh, 1990; Leary, 2003; Sedikides, 1993). The most powerful self-motive is the urge to feel good, have high opinion about oneself and to maintain high level of self-esteem. This motive is most often named self-enhancement (Krueger, 1998). Other motives are self-verification, which should be rather called "self-confirmation"[23] (Wojciszke, 2002), as this motive consists of leading to seek confirmation of a given self-image and an urge towards keeping a congruent view of oneself. Self-verification is in fact quest for the confirmation of our views (both positive and negative) about oneself (Sedikides & Gregg, 2008). These two motives are though responsible for keeping self-esteem high and buffering threatening stimuli and feedback from the environment. The other two motives, self-assessment and self-improvement, are geared towards performing a true diagnose of the self and towards improvement of modifiable traits in order to be better (Trope, 1986). Thus, these motives lead to better self-knowledge and improving in the areas that need so (Sedikides, 1993). In other words the motive behind self-enhancement is to perceive "I" as positive, behind self-verification to make "I" congruent, behind self-assessment to make "I" true to reality and behind self-improvement to make "I" better, more beneficial (Kossowska, 2009; Sedikides & Strube, 1997). Of course, the motives can compete with each other as they lead to different cognitive, affective and behavioural outcomes (Sedikides & Gregg, 2008). Thus, the two most powerful self-motives are not leading to a veracious self-image (self-description) but to an overly positive image that serves to reduce negative impact of everyday slings and arrows and to give an individual energy to perform usual life actions. Small wonder then, that a biased process of self-consciousness leads to formation of a biased structure of self-knowledge (Zaborowski, 1989).

Nonetheless, people are capable of truthful self-evaluation, but it requires mobilisation of cognitive resources. Many researchers claimed exactly so and pointed out that an accurate, diagnosing self-insight is oftentimes not only threatening for an individual but it is also effortful, e.g. Rogers (1961) claimed that obtaining more accurate self-knowledge requires first to overcome anxiety and to switch off defensive mechanisms. Similar viewpoint was presented by Birney, Burdick and Teevan (1969) who stated that anxiety reduces self-diagnose and that the "fear of failure" is a stronger affect than the

[23] This "truer" name, better reflecting the sense of this construct is reflected in the Polish term for this motive proposed by Wojciszke (2002): "samopotwierdzanie".

drive to have an accurate view of oneself. Obviously, the above-mentioned differentiation on central and peripheral domains of the self has certain importance also in case of self-enhancement as forming elevated self-assessment and ignoring negative feedback is much more probable for important than unimportant dimensions (Sedikides & Green, 2000). From the point of view of a survey methodology these are not great news- participants tend to yield superficial and *ad hoc* created assessments or opinions (even pseudo-opinions, Bishop et al., 1986) in peripheral domains, whereas in case of central dimensions they tend to present an elevated, overly positive image as spurred to do so by self-enhancement motive. This evidence points out, that apart from intended behaviours of inaccurate self-presentation (e.g. faking; Ziegler et al., 2012), there are habitual distortions of self-descriptions, that are not "special cases"[24] but they are rather a default, ordinary way in which self-knowledge is perceived and communicated. For sure, the communication rules characteristic for a given individual also play role in how answers are given in a survey (cf. Snyder, 1974 on self-monitoring) but it seems that self-enhancement (S-E) is a major force underpinning any self-description in any research context. Thus, it is warranted to analyse this self-motive more closely in order to gauge its possible implications for survey methodology in general and for response biases in particular. To this end, mechanisms of S-E need to be reviewed, its correlates examined and possible effects for self-response measurement scrutinised.

3.4 Self-enhancement: mechanisms, correlates and consequences

3.4.1 Definitions and terms

Due to large number of overlapping terms in the field it is warranted to clarify basic definitions and conceptions regarding self-enhancement in the first place. Sedikides and Gregg (2008) proposed that self-enhancement can be understood as: a) an observable effect (e.g. elevated scores in a measure, inflated self-ratings), b) an ongoing process (e.g. ascribing self-serving attributions), c) a personality trait ("repetitive inclination to demonstrate the motive") and d) an underlying motive (authors define it as a conscious motive to see oneself superior and strategically tailor social comparisons to prove one's superiority, e.g. by comparing oneself to weaker opponents; note, that this motive can also be non-deliberate). These biases are omnipresent and are part of the "false uniqueness" bias, a tendency to overclaim socially desirable traits to express the belief of being unique ("my children are the smartest, I am mostly touched by the tragedy, my problems are the most serious, I am the smartest worker", etc.; Chambers, 2008). This larger effect is nicely illustrated by an "emotion intensity bias" which describes a phenomenon to ascribe more intensive emotions to oneself and even to the members of one's own social group in comparison to others and people from other social groups (Chambers & Suls, 2007).

The effect is known under many different names such as rose-coloured glasses, self-enhancement, positive illusions, (overly) positive bias, self-serving bias, self-flattering, self-positivity, overconfidence, exaggerated positivity, automatic optimism, wishful thinking, unrealistic optimism, self-valorisation, self-ingratiation, self-insight failure, self-deception and above-average effects[25] (Beer, 2014; Beer & Harris, 2019; Kim, Chiu & Zou, 2010; Robins & Beer, 2001). Only a cursory glance at the above, much incomplete list[26] suffices to agree with Beer and Harris (2019) that the framework is plagued by "definitional issues". Moreover, the terms are not only plenty but they are also little defined (Alicke,

[24] As in SDR conceptions of self-enhancement or self-deceptive biases (Paulhus, 2002).

[25] The reversed effect, although much less common and much less researched alike, is also known. It is most often called self-effacement, self-derogation or self-diminishment (Kim, Chiu & Zou, 2010).

[26] For a similar lists of terms in Polish see e.g. Kossowska (2009) and Szpitalak (2012).

Sedikides & Zhang, 2019) contributing to the "bewildering array of phenomena that contain "self" as a predicate" (Leary & Tangney, 2003; see also Leary, 2004).

It is important to note that not all of the above terms are synonymous, like e.g. overconfidence has little to do with self-enhancement (Bensch et al., 2017) and e.g. above-average effect describes only one particular result (or a strategy) of the self-enhancement motive (Taylor & Brown, 1988). Other propositions entail creating one umbrella term to name all the similar or even identical effects. Bensch and colleagues proposed "positivity bias", whereas Cahill (2015) came up with "convenient bias". Another often used term is "motivated biases" (of self-perception; e.g. Chambers & Windschitl, 2004). The best term to describe these phenomena should denote their result, which is a self-image biased towards desirability and positivity, but also should refrain from implying exact processes lying underneath such over favourably self-descriptions (c.f. Cahill, 2015). In my opinion among the three above-mentioned terms the (overly) positive bias (of self-perception) is the least confounding and describes the construct best. Of course it is an overarching term and many of the above-listed terms can also be used to denote specific processes or effects, while others are only synonymic terms at best, or entirely confounding at worst.

3.4.2 Mechanisms

In many conceptions of human cognition accurate perception of the self and the surrounding environment is one of the most basic functions of cognitive system (Kelly, 1955). How it is then possible that such a common and omnipresent distortion of reality as the positivity bias could be present? What are the precise mechanisms driving it and what, if any, functions it serves ((Heine, Lehman, Markus & Kitayama, 1999)? These questions are also important from a more applied perspective, as no serious method to control overly positive self-reports can be conceived without the proper understanding of its mechanisms and the exact functions it plays in human cognition (Cahill, 2015).

The mechanisms suggested to stand behind the positivity bias are grouped into two main categories: motivational and non-motivational biases. The first group combines mechanisms that are driven by a motive, in other words, that certain self-relevant, evolutionary goal, achieving which is beneficial for an individual stands behind them (Sedikides & Gregg, 2008). In most of the cases the four basic self-motives (Bargh, 1990; Leary, 2003; Sedikides, 1993) are seen as processes underlying positivity bias in the motivational account. The second group entails mechanisms that are by-products of other processes and do not serve for any particular self-relevant goal (Beer, 2014; Cahill, 2015; Chambers, 2008). According to the motivational explanation, positivity bias serves for self-esteem protection from anxiety and threats of everyday life, some even proposed that self-enhancement offers defence against mortality (Pyszczynski, Greenberg, Arndt & Schimel, 2004). Other popular conceptions contended that positivity bias served as an energising principle (Sedikides & Skowronski, 2000), boosting individual's self-esteem and self-efficacy, thus contributing to one's goal realisation and effort expenditure (cf. Coopersmith, 1967). Yet another idea for a motivated positivity bias is that it helps to maintain self-esteem on a sufficiently high level to ensure positive social relations. This idea is based on the conception of self-esteem as an index of social value (sociometer) that helps to win partners and to associate with valuable social groups (Leary & Baumeister, 2000). It is worthy to note that motivational conceptions of self-enhancement point to its close relation to self-esteem and suggest that self-relevant goals are achieved through maintaining positive self-esteem. In that account positivity bias exists to protect and enhance self-regard when needed (Sedikides & Gregg, 2008). Arkin (1981) conveyed this notion by proposing two mechanisms of self-advancing and self-protecting: the former to boost the positivity of self-regard and the latter to diminish the negativities. This differentiation is in fact confirmed by research results as different strategies of self-enhancing are used by people with low *versus* high self-esteem (Kunda, 1999). Those with low or unstable self-esteem use

more self-protective strategies, while those of high self-esteem perform more self-advancing techniques. Kim, Chiu and Zou (2010) also pointed to these strategical differences: self-effacers used self-handicapping or effort withdrawal techniques that were less popular among self-enhancers. It is interesting to note that accurate self-assessors used other strategies, e.g. preparatory effort or remedial actions.

Arkin's (1981) differentiation on two types of techniques used to boost self-positivity, namely enhancing positive qualities and suppressing knowledge of negative ones, brings about another important notion in the science of overly positive bias: self-deception. Self-deception is by some seen as a type of self-enhancement but by others as a superordinate term to S-E; in this work it is treated as one of many manifestations of the overarching positivity bias (Hepper, Gramzow & Sedikides, 2010). Chance and Norton (2015) gathered three most frequent self-deception definitions: a) motivated false belief (Mele, 1997), b) motivated false belief in spite of disconfirming evidence (Greenwald, 1997; Sharot, Korn & Dolan, 2011), c) motivated and conscious false belief held simultaneously with a conflicting, unconscious and true belief (Sackeim & Gur, 1978; 1979). These definitions can be complemented by propositions brought by Snyder (1984) who defined self-deception as "motivation to tolerate discrepant self-image" and by Zaborowski (1989) who concentrated more on a processual side of the construct and proposed that self-deception is "selective avoidance of threatening information". Nevertheless, classical accounts on self-deception were criticised on the grounds that they "require that self-deceivers must hold two contradictory beliefs, while remaining unaware that one of them is held. This capacity has been deemed paradoxical, or even impossible, on logical, philosophical, and psychological grounds" (Peterson et al., 2003). Mele argued further that intentionality and consciousness is not needed to speak about a "motivational bias" (1997). Also Greenwald (1980, 1997), Mele (1997) and Peterson (1999) do not agree with the traditional definitions (e.g. Sackeim and Gur's) and point out that self-deception is not based on objective, but rather subjective, affectively-marked premises, as person may ignore conflicting evidence unwillingly, e.g. due to lack of resources to process it properly, due to lack of knowledge, intelligence, experience, etc. This line of thinking was picked up by von Hippel and Trivers (2011) who proposed a new opening for the self-deception research. Their definition resembles this of Zaborowski (1989) and contends that self-deception is simply a process of biasing the information-gathering process in a way favourable to an individual. What is currently favourable for a given individual is decided by his own goals and motives. The debate on the functions of self-deception is still ongoing, with most of the researchers agreeing that it serves adaptive purposes like: a) deceiving others, b) gaining social rewards and c) reaping psychological benefits (Cahill, 2015; Chance & Norton, 2015). Some voices also point that self-deception may be also maladaptive (e.g. Peterson et al., 2003)[27].

Thus, the traditional "motivationalists" saw positivity biases as serving to protect self-esteem, offer comfort and provide self-satisfaction (Taylor and Brown, 1988). The ideas of von Hippel and Trivers (2011) changed the focus from internal effects of self-deception (self-esteem protection) to external functions, namely enhancing social status. Keeping self-regard positive may also serve to motivate oneself and mobilise resources to sustain effort and achieve one's goals (Coopersmith, 1967; Sedikides & Skowronski, 2000). The recent evidence presented by Gesiarz, Cahill and Shalot (2019) and Kappes and Sharot (2019) suggests that positivity bias is in fact motivated, but that it is also more automatic and working on earlier stages of information processing than it was previously thought. In the study by Gesiarz and collaborators (2019) the participants were playing a game in which certain responses were more desirable than others as leading to higher monetary gains. However, the game was promoting only accuracy and participants could not gain money from responding desirably but

[27] Please see the discussion in part 3.5.3 for a more detailed elaboration on adjustment of positivity bias.

incorrectly. Despite that the participants already from the early beginning of the task presented a bias towards choosing desirable responses. Moreover, this bias increased further during the game. The authors point to the bias in responding but also in the processing of the incoming information in the way that desirable responding was faster and came after analysing smaller amount of information than undesirable one. Hence, the desirable conclusions are made on the basis of smaller evidence and with little effort to verify their accuracy. These results show that the positivity bias is motivated by participants' expectations but it is also automatic (non-deliberate), and active already on the very early stages of cognition. Moreover, the results obtained by Kappes and Sharot (2019) showed that cognitive resources are not needed to yield cognitive bias, thus corroborating the automatic character of positivity bias formation. It is not known however, whether the inclination towards bias is learnt (e.g. in the process of socialisation), innate or "a combination of both".

On the other hand, the nonmotivational accounts of self-enhancement simply see limited cognitive resources (information processing limitations) as the main mechanism of this bias (Beer, 2014; Chambers & Windschitl, 2004). Beer, Rigney and Koski (2018) suggested that self-enhancement can be motivated and orchestrated by focusing on positive aspects of oneself, selecting only positive feedback and selecting only social comparisons favourable to self, e.g. by choosing inferior referents or selecting only these dimensions in which an individual was good at. Similarly, Cahill (2015) sees the basis of enhanced self-descriptions in biased belief updating, e.g. in discarding negative feedback or biased selectivity of relevant/irrelevant feedback. Cahill (2015) also contends that non-motivated, automatic processes are sufficient for positivity bias, moreover, he thinks that controlled processes seem to serve for bias correction and not its evocation. There is also a question whether the negative feedback is simply not encoded or is it suppressed at recognition (bias at encoding or at recognition phase)? Study 2 in Rigney's research (2019) shows that even if the incentives to claim negative feedback occur before encoding they do not alleviate the positivity bias which was attenuated only in the self-irrelevant-high incentive condition. However, Study 3 in the same research that used event-related potentials (ERP) measures seems to show that the negative feedback is not "forgotten" nor concealed, it is simply not encoded in the first place (Rigney, 2019). This finding is somewhat corroborated by another ERP study that shows that goal-irrelevant feedback has reduced processing in comparison to the goal-relevant one (Severo, Paul, Walentowska, Moors & Pourtois, 2020). These results pushed Rigney (2019) to postulate two mechanisms of positivity bias: a) different standards for positive and negative feedback- the negative one is simply not encoded, b) depth of processing when searching through memory- the search is deeper when positively-evaluated object is to be evaluated positively- e.g. if we like a politician and we have to decide whether she is intelligent, then the memory search for any evidence to confirm her intelligence will be much more thorough in comparison to assessing intelligence of someone we do not like. Both these mechanisms point to positivity bias as very basic and automatic process[28].

The account of self-enhancement as an automatic process also suggests that self-evaluations' aim is to claim one's positivity, hence reduced processing of negative feedback due to its goal-irrelevance- as the goal is to show oneself in a positive light, everything that impairs this process is discarded on early stages of cognitive processing. Moreover, it also suggests that positive self-esteem, characteristic for a vast majority of people (e.g. Różycka & Wojciszke, 2010), has its consequences for self-evaluations- as the self-esteem is mostly high it means that in general people like themselves, which entails looking hard for any piece of evidence confirming good characteristics. In the line of this reasoning, Swann and Read (1981a, 1981b) showed that people actively seek to confirm self-schemata and that feedback

[28] Epley and Whitchurch (2008) showed an example where self-enhancement is even surprisingly automatic: in their research one's face was recognised faster when it was made more attractive (enhanced) than when it is made less attractive.

that confirms them is valued as more valid than feedback that negates them. In the terms of mechanisms suggested by Kozielecki (1981) most of the people seem to use the more lenient alternative method when assessing themselves, resulting in overly favourable self-images. These findings are confirmed also by other research, e.g. Peterson and colleagues (2003) showed that people with high SDR tendencies (as measured by the BIDR or the Eysenck Lie Scale) ignore negative outcomes of gambling games and proceed in such tasks in a perseverative way. In other study, Peterson, Driver-Linn and DeYoung (2002) showed that high self-deceivers were slower in identifying anomalous visual stimuli. In line of this proposition, Chance, Norton, Gino and Ariely (2011) showed that self-deceptive participants failed to recognise what led them to achieving good results in a task (researcher's help). They have, hence, committed an attributional error and ascribed their success to their own abilities. Moreover, it pushed them also to predict similarly favourable outcomes in the future, even if they were financially motivated to predict accurately. All these pieces of evidence point to lower error utilisation and larger rigidness of behaviour in self-deceiving individuals[29].

Other possible non-motivated mechanisms of positivity bias were proposed by Chambers and Windschitl (2004) who differentiated between upper tier (general, broad mechanisms) and lower tier (specific) processes. The upper ties processes are proposed to operate at all three stages of responding (information recruitment, evaluation, judgement formation) whereas the lower tier mechanisms are to operate only on one of them. The first group entails egocentrism, focalism and generalised-group account, whereas the second counts in differential accessibility, differential attention, case vs. rate assessment, idiosyncratic standards, differential standards, regression-to-the-mean (small role), differential confidence (small role), anchoring and insufficient adjustment. Chambers and Windschitl (2004) proposed these mechanisms as main non-motivational processes standing behind the observed positivity bias, mainly in tasks requiring comparative judgements, e.g. "What are math abilities in comparison to an average person from your class/school?" It is unclear how these mechanisms would perform in case of items without any direct comparisons required, e.g. "What are your math abilities?"

Thus, the "non-motivationalists" also have succeeded in gathering evidence to support their view on the mechanisms of positivity bias. However, it seems that none of the proposed mechanisms can account for all of the findings, as in example the results showing that threat to self-esteem increases bias, whereas affirmation reduces it (Paulhus et al., 2003; Gramzow & Willard, 2006; Kumashiro & Sedikides, 2005; Trope & Neter, 1994) go against the non-motivational approach. The results collected by Chance and collaborators (2011) in a study when social recognition (participants were praised for good results in a sham test) exacerbated self-deceptive tendencies can serve as a confirmation for that.

Most probably there is an interplay between motivated and non-motivated processes and both accounts can be true in certain conditions or for certain participants (Chambers, 2008; Moore, 2005). This is suggested by results that affirmation does not reduce positivity bias when it is caused by cognitive mechanisms, e.g. memory bias, and works only when bias is motivational (Gramzow & Willard, 2006). On the other hand evidence that positivity bias is increased under cognitive load (cognitive load forces more automatic processing due to lack of cognitive resources to manage all of the actions) points to the non-motivational account (Paulhus, Graf & van Selst, 1989).

In line with the above findings it was also showed that cognitive load increased self-flattering ("above average") ratings. Moreover, the response times were three times faster under the cognitive load condition, indicating that under cognitive load stimuli processing is automatized, resulting in fast but also self-enhancing responses. This suggests that the above-average overly positive self-judgments are

[29] These effects are probably also somewhat related to need for cognitive closure that can also be described in the self-motives framework.

due to heuristic processing of information, e.g. due to decreased adjustment processes or lower accessibility of relevant self-knowledge (Paulhus et al., 1989; Paulhus & Levitt, 1987). It also shows that an accurate self-rating is a controlled process that needs cognitive resources[30] (Kruger, 1999).

3.4.3 Neurocognitive and physiological evidence on positivity bias

This double explanation is corroborated by the research evidence from neurocognitive studies pointing that the neural mechanisms of exaggerated positivity differ depending on whether self-esteem is *versus* is not threatened (Beer, 2014; Beer & Harris, 2019; Hughes & Beer, 2012). In general, medial OCF (orbitofrontal cortex) and dorsal ACC (anterior cingulate cortex) activation is related to self-rating accuracy (overcoming positivity bias) and to reduced self-flattering judgements, whereas MPFC (medial prefrontal cortex) activation is related to egoistic positive bias (Beer & Hughes, 2011). This structure (MPFC) also plays important role in constant self-evaluation (Beer, 2007). Additionally, ventral ACC (vACC) is related to detection of valence (positive *versus* negative) and it may also encode social desirability of traits, coding what is desirable or potentially rewarding (ventral ACC also play role in expecting/maximising reward in gambling). These results suggest that high quality data (non-biased) can be expected from participants that yielded high OFC and dorsal ACC activation during a task. However, the research presented by Beer and Hughes (2011) and Flagan and Beer (2013) suggested that the neuronal correlates of self-positivity triggered by self-esteem threat *versus* those caused by cognitive load are different. The former are related to an *increased* activation of medial OFC, ventral ACC and MPFC with a concurrent activation of a wider neural network involving also basal ganglia structures (see Table 1 in Flagan and Beer (2013) for a precise list of areas), whereas the latter are linked with a *decreased* activation in medial OFC and other structures in the frontal lobe. It is also worthy to note that when self-esteem threat is present it does not matter whether the cognitive resources are at disposal or not- self-enhancement takes place nonetheless (Beer & Hughes, 2011).

Another evidence for automaticity of self-enhancement was brought from an EEG research by Krusemark, Campbell and Clementz (2008) where ERP source and voltage analysis was used to show that unbiased attributions were preceded by increased dorsomedial frontal cortex activation, meaning that greater cognitive control was needed to prepare them in contrast to biased attributions. Similar results from an fMRI study were provided by Hughes and Beer (2012), where accountability condition in a self-descriptory task led to increased activation of dorsal ACC, OFC and MPFC resulting in less self-serving bias in comparison to non-accountability condition[31]. Additional information was brought by Kwan and colleagues (2007) who conducted a transcranial magnetic stimulation (TMS) study: when TMS was delivered to MPFC and precuneus it resulted with less self-enhancement in comparison to conditions where sham TMS was used or when TMS was delivered to other brain structures, not related to cognitive control, e.g. supplementary motor area (SMA). The authors pointed out that their results corroborated the role of MPFC in self- and other-deception. Almost exactly the same results from a similar study design were obtained by Amati, Oh, Kwan, Jordan & Keenan (2010), whereas yet another TMS study conducted by Barrios and others (2008) showed that MPFC is important for agentic self-enhancement but not for moralistic one. Why and Huang (2011) showed that self-enhancing opinions about one's performance in demanding cognitive tasks reduced cardiovascular responses, namely, self-enhancing participants had reduced heart rate as well as systolic and diastolic blood pressure. Moreover, some of the self-enhancement effect was mediated by perceived control over the task that was gained during the first phase of the task. The authors contended that self-enhancement

[30] As all effortful processes are impaired under cognitive load.

[31] Hughes & Beer (2012) used a customized version of the OCT. Higher c parameter was taken by them as an indicator of less self-serving responding.

may be related to short-term stress coping reaction that may be beneficial for health. Why and Huang's results and conclusions were also confirmed by Hernandez and co-workers (2015).

Hence, neurocognitive and psychophysiological studies have brought very interesting results giving a unique insight into mechanisms of self-enhancement. Of course, many of these investigations suffer from common limitations of neuro studies, like small number of participants and simplified experimental stimuli. A good example comes from the study by Kwan and colleagues (2007) where only twelve participants took part and the self-enhancement measure used was very crude ("Does this adjective describes you? Yes or No?").

3.5 Positivity bias: state- or trait-driven?

One of the classical debates in social sciences is about the temporal nature of concepts (whether they are rather stable or are they rather transient) as well as their inter-individual variability (whether they change or not across people). Such state versus trait debate is also present in the self-enhancement framework. Taylor and Armor (1996) as well as Sedikides, Herbst, Hardin and Dardis (2002) emphasised the sensitivity of self-presentation strategies to "situational demands and social settings", hence siding with the "state" account. The opposite view was presented by Vecchione, Alessandri, Barbaranelli and Caprara (2013) who contended that self-enhancement is a trait-like characteristic that is stable in time ("trait" account). Paulhus and Trapnell (2008) also sided with the trait "team", showing that self-enhancing is related to the trait-like attunement to self-presentation demands and to the nature of the image intended to be presented (agentic or communal). The researchers considered that in case of certain individuals the tendency to self-enhance is always present, but the exact content of their self-presentation is context-dependent. For example when there is a need (motive) to avoid social disapproval a moralistic (communal) self-presentation may be deployed, whereas when there is a need (motive) to protect self-esteem an agentic (egoistic) self-presentation will be used. Fitting self-presentation to the specific audience and situation would point to Ziegler's (2015) conception that positivity bias is also the result of an interaction between trait and situational factors. These account was also confirmed by some research results, e.g. by Robins and Beer (2001) or by Paulhus and Harms (2004) who showed that results of the overclaiming questionnaire (OCQ) respond to both state and trait engendered self-enhancement.

However, as self-enhancement is different in people with low versus high self-esteem it is believed that some differences exist between egoistic and moralistic positivity biases. Vecchione and Alessandri (2013b) suggested that egoistic are more stable, whereas moralistic more transient, but also concluded that both egoistic and moralistic self-enhancement have more trait than state variance, though both have less stability than Big Five personality traits. Similar results were provided by Schmitt and Steyer (1993) who in a multi-trait, multi-technique (method) study (MSMT) presented that egoistic bias correlated more with the trait variance of emotional stability and conscientiousness, whereas moralistic bias with the state variance of these variables. Interestingly, Vecchione and Alessandri (2013a) showed that moralistic and agentic self-enhancement were not correlated with each other (the correlation disappeared when controlled for method variance). This result is in accordance with the original Wiggin's (1964) theory of two orthogonal SDR factors: alpha and gamma. It is also in line with the theory of two unrelated community and agency motives (Trapnell & Paulhus, 2012). The separateness of these constructs is also suggested by their different neuronal basis (Barrios, Kwan, Ganis, Gorman, Romanowski & Keenan, 2008; Farrow, Burgess, Wilkinson & Hunter, 2015). These results led Vecchione and Alessandri (2013a) to define self-enhancement more in the framework of a personality characteristic, than response bias. This line of reasoning seems justified as John and Robins (1994) pointed to sizeable individual differences in the validity of self-assessment: from great self-

enhancement to sizeable self-effacement. If and how it is related to the rules of generating self-judgements described by Kozielecki (1981) is yet to be determined.

Researchers attempted to find explanations for individual differences in positivity bias in line with the trait account and proposed that a) cognitive biases (Klar & Giladi, 1997; 1999), b) lack of metacognitive ability (Kruger & Dunning, 1999) or c) lack of motivation to self-enhance (Sedikides & Strube, 1997) can play deciding role in explaining inter-individual variance in tendencies to self-enhance. Kim, Chiu and Zou (2010) suggested that this may depend on the domain assessed as new domains may be more based on the metacognitive processes, whereas well-known domains may rely more on motivated processes. However, no controlled comparisons of this statement were ever presented. Chance, Gino, Norton and Ariely (2015) provided another evidence supporting trait account in the state vs. trait discussion. They have pointed out that the slow decay and quick revival of self-deception suggested trait-like explanation of the phenomenon. They have shown that self-deceptive judgments are very stubborn to go away, even in the face of direct evidence against them. Self-deceivers, even confronted with reality, fail to learn not to self-deceive- they are susceptible to succumb to self-deception on another occasion which, according to Chance and colleagues, implies that personal characteristics are more important. On the other hand, there is also evidence supporting rather state account, e.g. that bias can be attenuated by practice (Kruger & Dunning, 1999) or by monetary incentives to report accurately (Kim et al., 2010; but see Gesiarz et al., 2019).

However, the research by Luo, Sedikides and Cai (2019) brought evidence about substantive heritability of self-enhancement manifestations, such as narcissism levels, self-favouring judgments and overclaiming knowledge, thus scoring one point for the "trait" team again. It is not known as yet, whether this heritability is related to self-esteem heritability established by Saphire-Bernstein and collaborators (2011). Due to the known links between narcissism, self-enhancement and self-esteem (Hepper, Gramzow & Sedikides, 2010) such analysis would truly show what traits and tendencies are heritable[32].

3.5.1 Context- and individual-related correlates of positivity bias

Accountability

Measurement context and individual correlates of positivity bias were also researched in order to deepen the knowledge about this phenomenon. One of the characteristics of the measurement situation manipulated was accountability which seems to lower self-enhancement (Lerner & Tetlock, 1999; Sedikides, Herbst, Hardin & Hardin, 2002). Beer and collaborators (2013) suggested that social accountability may be especially effective in curtailing the exaggerated positivity because it involves losing social face (huge threat to self-esteem) in case of self-enhancement, other accountability types (e.g. losing financial remuneration) may not be self-esteem threatening and, thus, not that effective in reducing self-enhancement. On the other hand, Niessen, Meijer and Tendeiro (2017) showed that self-reports from "applicant" contexts have lower predictive validity than self-reports from "research" contexts (the hypothesis is that self-enhancement in applicant conditions lowers the validity), moreover, the self-reports burdened by self-enhancement did not offer incremental validity over other measures, whereas those not distorted did offer an R^2 increase when added to cognitive tests and graded-point average (GPA) in a regression equation.

[32] See also research of Brummelman et al. (2015) linking narcissism with internalization of parents' inflated views of their children in the way of social learning.

Item and domain characteristics

Certain role may be also played by the type of items used. It is worth to note that *direct* questions (inducing comparison, e.g. "How good at math are you in comparison to an average pupil?") yield larger enhancement effects, e.g. above-average effect than *indirect* questions ("How good at math are you?"), which Chambers and Windschitl (2004) attributed to a number of non-motivated cognitive biases[33]. They also claimed that many of the nonmotivated biases do not apply to indirect questions. Hence, the most popular format of self-report items in the social sciences, the rating scales, e.g. of the Likert type, are "safe" from these distortions as they are indirect questions that do not request any comparisons.

Interesting results were brought by analyses of domain characteristics that confirmed that people self-enhance mainly on things that are important, central to them, according to the general rule: "self-centrality breeds self-enhancement" (Gebauer et al., 2012; Gebauer, Wagner, Sedikides & Neberich, 2013). An attempt to systematise the findings showed that: a) trait (domain) social desirability, b) domain familiarity, c) certainty and d) trait importance (Crocker, 2002; Sedikides et al., 2003; Hughes & Beer, 2012) all predicted magnitude of self-enhancement bias. It is thus evidenced that if the trait is desirable for a given participant she will self-enhance more in items related to it (Phillips & Clancy, 1972). Other evidence on this topic was brought by Gramzow and Willard (2006) and Paulhus and colleagues (2003) who also claimed that more socially desirable and rewarding traits are more exaggerated. On the other hand, when the domain is not familiar, people yield more accurate ratings (Beer, Lombardo & Bhanji, 2010). Sedikides and Strube (1997) ascribed this latter effect to larger tendencies to self-assess than to self-verify or self-enhance in situations of low certainty which may relate to the fact that people self-enhance more in domains that are hard to verify. Objective verification seems to reduce self-enhancement, so people mostly enhance in domains hard to verify, e.g. less public, less specific and less objective. Beer and Hughes (2010) differentiated between "broad" and "narrow" traits with self-enhancement more related to the "broad" ones, namely traits which are simply less verifiable than others as their definitions are more inter-individually varied, e.g. creative versus tidy; "tidy" is easy to define, whereas "creative" is more difficult and have more varied definitions among people. Van Lange and Sedikides (1998) claimed that self-enhancement is augmented when subjective and abstract traits are rated, in comparison to traits that are objective and verifiable. Also Felson (1981) showed similar effect with college American football players who over-rated more on "ambiguous" versus "unambiguous" traits. Allison, Messick and Goethals (1989) called this a "Muhammad Ali effect", who said "I only said I'm the greatest, not the smartest". Obviously one's intelligence is much easier to gauge than one's greatness. Not surprisingly, people tend to value more the domains they are good at, which complicates the issue even further. Hill, Smith and Lewicki (1989) showed that students high on certain school skills valued them as more important and more desirable than lower achieving students did. What is more, the perception can and often does change, as e.g. failure in a given domain leads to perceiving it as less important than before the failure (Hill et al., 1989). This view was also expressed by Tesser (1988) in the self-evaluation maintenance theory, where three factors were deemed critical: a) performance in a domain, b) relevance of that domain and c) its relationship to target; all related positively with self-enhancement tendencies. Moreover, also task difficulty can influence self-enhancement processes. As evidenced by Kruger (1999) people tend to overstate when assessing easy tasks and tend to understate when assessing difficult tasks. Moore and Cain (2004) even concluded that this effect can account for greater

[33] Egocentrism, focalism, anchoring, case account, regression-to-the-mean account, hybrid referent group, generalised group effect, etc.

rates of entrepreneurial entries into industries that are perceived as "easy"[34], e.g. gastronomy, in comparison to businesses deemed difficult, e.g. bridge construction. Thus in a survey on academic or professional skills participants should have a general tendency to understate (efface) their abilities in the domains deemed difficult, e.g. math, and overstate (enhance) them in the domains seen as easy, e.g. foreign languages (see Baczko-Dombi, 2017 on students' opinions on school subjects).

To sum up, the research evidence shows that different traits can be differently biased, due to their importance, desirability, verifiability and controllability, as perceived by participants (Alicke, 1985; Dunning et al., 2004; Gebauer et al., 2012; McLarres & Oyelere, 1999; VanYperen, 1992). This is corroborated by Leary's summary on conditions when overly positive self-presentations- such take place only when: a) the situation is important for a person, b) it is unlikely to be caught on an inaccurate self-presentation and c) untrue self-presentation will not harm anyone (Leary, 1999).

Personality

Self-enhancement tendencies also correlate with a number of individual differences. First of all, they are linked with types of personality that are known for such propensities: narcissism and self-monitoring. Narcissism is thought to be a basis for self-enhancement tendencies that "cannot be bridled" by measurement context, stakes, accountability or instructions (Robins & John, 1997), whereas self-monitors self-enhance when it is opportunistic and can bring gain for them. Hence, narcissists are believed to self-enhance always, on the other hand, cautious impression management of self-monitors is strictly related to the measurement context and stakes (Paulhus, 2003). Further research revealed a connection between self-enhancement and narcissistic tendencies, ego-involvement in the task, self-serving attributions of performance e.g. self-enhancers see the root of their success, but not the failure in their abilities, failure is always attributed to external phenomena. Moreover, self-enhancers tend to attribute success to their underlying ability, not their motivation (Quattrone & Tversky, 1984; Fernbach, Hagmayer & Sloman, 2014). This bias is called "affect regulation in response to the threat of failure" and it is believed to protect fragile and unstable self-esteem (Robins & Beer, 2001). The results provided by Trope and Neter (1994) confirmed this pattern as participants in negative mood or after a negative feedback yielded more self-enhancing responses than when in good mood or after a positive feedback. This finding confirms conclusions that self-enhancement is increased in (ego-)threatening situations. In another study Robins and John (1997) showed a group of narcissists a video recording with their performance. The presentation led to increased instead of decreased self-enhancement, which again point to the existence of processes that actively distort the feedback (cf. Rigney, 2019).

More basic personality traits were also correlated with self-enhancement. In an analysis of the Polish post-PIAAC study Rynko and Palczyńska (2018) achieved positive correlations between the Grit scale and overstating one's ICT skills, although no correlations were obtained between the self-assessment of these skills and the Big5 personality traits. In fact, self-enhancement *should* correlate with neuroticism and conscientiousness as those two personality traits correlate with self-handicapping- one of the most important behavioural strategies used to protect self-esteem (Ross, Canada & Rausch, 2002). Vecchione and Alessandri (2013b) indeed found evidence for such correlations. However, these relations are at best very modest in magnitude, as shown by Tonković, Galić and Jerneić (2011) who found a very small negative correlation between OCT bias and neuroticism- -0.11. Personality traits also show a rather stable pattern of relations with positivity bias as measured by SDR scales. Positive correlations between SDR scores and extraversion, agreeableness, conscientiousness and openness to

[34] This effect probably also extends to other spheres of human activity, e.g. on electing college careers; this effect could be especially prominent in countries where higher education is free of charge, e.g. Poland.

experience were reported, whereas negative relation was established with emotional stability (neuroticism; Paulhus & John, 1998; Pauls & Stemmler, 2003). Obviously these correlations can be interpreted in two ways: that an individual yielding socially desirable responses is in fact a person with this (very desirable) pattern of personality or that both measures, SDR scale and personality scale, were distorted in the same direction by self-enhancing tendencies (McCrae & Costa, 1983). So far, this research on disentangling the substance *versus* style dispute seems far from delivering final answers (Ziegler, 2015).

Intelligence, educational level and cognitive abilities

More research evidence is available on the relation between self-enhancement and participants' intelligence and education. Both these variables can influence not only the results (enhanced self-reports), but also the process leading to distorted results as claimed by Christiansen, Burns and Montgomery (2005)- more intelligent (high IQ) participants can fake self-reports in more subtle ways, more difficult to identify. There is also evidence showing that the elderly and respondents of lower educational level are more prone to yield socially desirable responses (Fraboni & Cooper, 1989). Similar relations are observed also in the case of socio-economic status (SES), where responders of lower status are characterised with more SDR (Hrgović & Hromatko, 2019; Ross & Mirowsky, 1984), moreover this pattern is also typical for the countries of lower GDP *per capita* (Buckley, 2009; van de Vijver & He, 2014). These results may point to very interesting but unfortunately rarely explored evidence that SDR is related to automatic processing and low levels of cognitive abilities as suggested by research of Paulhus and colleagues (1989) where attentional load increased the proportion and speed of socially desirable responses.

Probably the best known account on the relation between cognitive abilities and adequacy of self-reports is the Kruger-Dunning effect (1999) which entails that participants with lower cognitive abilities yield less adequate (overly positive) self-assessments of one's traits that is warranted by more objective criteria. Kruger and Dunning (1999) explain this effect by linking lower cognitive abilities (e.g. lower IQ) with insufficient metacognitive abilities to appraise oneself validly which results in more self-enhancement.

Overall, there is large evidence that participants of lower education and/or lower cognitive skills yield responses of lower quality, be it biased in a socially desirable (overly positive) way, distorted by response styles or careless responding or simply sensitive to questionnaire alterations or item wording (e.g. Converse & Presser, 1986; Grau et al., 2019; Gummer, Roßmann & Silber, 2018; Jelonek, Worek, Turek & Muszyński, 2019; Krosnick, 1991; Meisenberg & Williams, 2008; Naemi, Beal & Payne, 2009; Paulhus & John, 1998; Rhodewalt & Morf, 1995; Roßmann et al., 2017; Van Vaerenbergh & Thomas, 2013). However, it is yet to be established what is the exact mechanism standing behind this effect. Malhotra (2008) pointed that it is not the main effect of lower cognitive skills (as proxied by lower educational level completed) but an interaction between educational level reached and response time: among the participants with lower education only those with fast response times yielded lower quality data. Other researchers, e.g. Krosnick (1991; 1999) and Tett and Simonet (2011), also inclined towards interactional explanation, namely that an interaction between ability and motivation, or even between opportunity, ability and motivation, can explain the quality of participant's responses. Roßmann and co-workers (2017) provided empirical evidence that also survey design affects participants' abilities and motivation to yield accurate responses: burdening designs may result in lower answers quality and higher measurement errors (Groves & Lyberg, 2010).

It is worthy to point out that these results and suggestions correspond very well with the Turner's (1975) and Kozielecki's (1981) ideas that creation of adequate self-knowledge requires significant

cognitive effort and skill. However, it is still not sufficiently known what is the exact mechanism that causes this pattern of results. Jackson and Messick (1958) attributed one of the response biases (acquiescence response style; ARS) to "low level of cognitive energy across situations", while initial research with the use of anchoring vignettes seemed to suggest that biased or inconclusive results stemmed from lower comprehension of items among lower educated participants, especially those below 12 years of formal education attainment (Bago d'Uva, van Doorslaer, Lindeboom & O'Donnell, 2008; Murray, Ozaltin, Tandon, Salomon, Sadana & Chatterji, 2003)[35].

On the other hand, Rynko and Palczyńska (2018) showed that higher educated participants self-enhanced more in the Polish post-PIAAC study. The same was true for people from big cities, in comparison to people from smaller locations. As the domain assessed were ICT skills it is probable that it is not the effect of education/IQ on self-enhancement but probably a matter of the domain- clearly ICT skills are more important for better educated and living in bigger cities where such skills are more valued on the labour market and where most IT companies have their offices. ICT skills overstating decreased with participants' age in this study, which seems to corroborate the "domain desirability" hypothesis (Alicke, 1985) over the "cognitive skills" theory (Kruger & Dunning, 1999)- ICT skills are obviously more important for younger Poles and there is a huge generational gap in ICT skills level in Poland (younger are more skilled; Burski et al., 2013).

Hence, the debate on the relation between cognitive abilities (as proxied by IQ or educational level) and self-enhancement techniques is still far from conclusions as even of the most central effects for this account, e.g. the Kruger-Dunning effect, was recently subjected to a heavy critique, mainly on methodological grounds (Gignac & Zajenkowski, 2020; He & Cote, 2019; Humberg et al., 2019). The caveats mainly regarded using difference or residual scores as main dependent variables in the original study which measures were proved to conceal, distort and mismatch information provided in the original measures (He & Cote, 2019; Humberg et al., 2019; Nuhfer et al., 2016; 2017; see also Zumbo, 1999). Other mathematical approaches to calculate such effects were advocated, e.g. polynomial regression and research surface analysis (Edwards, 1994; Edwards & Parry, 1993). The analyses performed by McIntosh, Fowler, Lyu and Della Sala (2019) also showed that metacognitive processes are "neither necessary nor sufficient" to cause the Kruger-Dunning effect. Krajc and Ortmann (2008) also criticised previous studies on the grounds of recruiting samples with skewed abilities (mainly students from very selective universities) that contributed to the alleged effect (see also polemic with these claims in Schlosser, Johnson, Dunning and Kruger, 2013).

Gender differences

Another important debate is around gender differences. In general, the relation of gender and positivity bias is not a very pronounced one, however, there are some systematic patterns observed. The research generally points that self-enhancement is used more often by men than women (Guadagno & Cialdini, 2007) and that also cultural norms in general see self-enhancement as a more masculine thing (Rudman, 1998). What is more, self-enhancing women are often perceived as less attractive (Sedikides et al., 2015). This is corroborated by the evidence that men enhance their competences and assertiveness (agentic traits) and women rather promote their social skills (communal traits) (Leary, 1995). This would suggest that men should be more prone to agentic (egoistic) bias than women and women should yield larger communal bias than men (Paulhus & John, 1998). However, in the analysis of Rynko and Palczyńska (2018) women indeed overstate slightly less than male, but this relation held only in math-related tasks (an agentic trait) and there were no

[35] However, this result needs further research, especially in the light of evidence criticising the methodology of anchoring vignettes (Stankov, Lee & von Davier, 2017).

differences in other, also agentic tasks. Hence, it can be concluded that group and cultural norms are important for across gender self-enhancing tendencies. These results also bring yet another confirmation on the key role of domain and items importance and desirability.

Agency and communion

The individual differences in self-enhancement were also analysed in the agency-communion framework. Paulhus and Trapnell (2008) integrated these dimensions with the SDR research, proposing two "self-presentational styles": one agentic in nature, based on claiming "superhero" abilities and related to self-enhancement, other communal in nature, consisted of ascribing "saint-like" qualities and related to impression management. Djikic, Peterson and Zelazo (2005) also related self-enhancement to egoistic bias (see also Paulhus, 1991 and Paulhus & John, 1998 for similar conclusions). In their research egoistic self-enhancers (people scoring high in the SDE subscale of the BIDR) showed a memory bias- they did not recognise negative feedback and in consequence noted more negative misses and less negative false alarms with rising SDE score (cf. Rigney, 2019). Participants with moralistic bias (high scores in the IM and the SDD subscales of the BIDR) spent more time viewing feedback, regardless its positivity. Djikic and collaborators (2005) ascribed this effect to early memory bias, that hinders processing of negative feedback. However, the correlations between the BIDR subscales and behavioural measures (e.g. response times) used by Djikic and colleagues only amounted to around 0.20. Moreover, Rigney (2019) criticised memory indices used in this research, questioning their validity as measures of memory bias. Using a modified procedure and a Signal Detection Theory (SDT) measure of positivity bias Rigney (2019) showed that positive traits are always more liberally recognised than negative ones, yielding higher accuracy for negative traits (*d'* index from SDT). Financial incentives and self-relevance of the traits did not change this pattern, indicating automaticity of positivity bias stemming from different "standards" for positive and negative feedback with the positive one more liberally accepted, regardless the conditions. Similar results, this time showing no positive effects of accountability and financial incentives on the accuracy of self-reports was shown by Ehrlinger, Johnson, Banner, Dunning and Kruger (2008). The results reported by Schlosser and collaborators (2013) showed that learning process also did not have any effect on the accuracy of self-reports, as surveyed students failed to increase their accuracy in the course of their education.

Another discrepant result was brought by Musch (2003) who showed results where memory (hindsight) bias was not correlated with egoistic bias, but with moralistic bias instead, as indicated by the German version of the BIDR. In another study Djikic, Chan and Peterson (2007) confirmed that moralistic self-enhancers did not yield this early memory bias, while egoistic self-enhancers did. However, they have shown that the bias was reduced in the latter group in the "social facilitation" condition (participants were recorded on a video camera when processing the feedback). The exact mechanism of this reduction is disputable though. The likely candidates are: a) accountability (someone may check the discrepancy between answers and recording, hence exaggerated self-enhancement may be identified), b) other self-presentation behaviour was sparked by the camera (camera as "the audience"), c) greater self-awareness was caused by camera recording, thus reducing the "habitual", biased way of processing. Note, that all of these explanations fit with the list of conditions that bridle self-presentational tendencies (Leary, 1999).

An important notion was raised by Abele and Wojciszke (2007) and Wojciszke, Baryła, Parzuchowski, Szymków and Abele (2011) that agentic values are more important to self-esteem than community values. However, there is a possibility that it is true only for Western, individualistic cultures. Kwan and colleagues (2009) obtained evidence for that as they showed that smaller percent of variance was accounted for by benevolence, merit and bias in self-esteem measure in Chinese sample (~30%) than

in American one (~70%), pointing to different factors constituting self-esteem in individualistic and collectivistic cultures. The cultural differences are beyond the scope of this work but it is worthy to note that there is large evidence supporting that agentic traits should be self-enhanced more than communal ones due to larger role of agentic traits for self-esteem maintenance, at least in individualistic cultures.

Conclusion

The evidence of individual differences in self-enhancement tendencies is mixed. There is evidence that these tendencies correlate with a wide array of traits, but on the other hand the correlations obtained are often very small and the evidence is inconclusive. On this level of research it seems that there is more evidence in favour of contextual than trait explanations of self-enhancement. One of the important factors affecting overly positive self-ratings is the domain being assessed, with domains important for participant, more subjective and less verifiable being more prone to bias. There is also evidence that accountability and inducing self-aware states is related to decreased positivity bias. There is also an ongoing debate on whether positivity bias is domain-general *versus* domain-specific. No definite solutions of this debate are possible now due to still insufficient evidence. It is worthy to note that it goes on from an early research period as already Brandt (1958) noticed that across-domain adequacy of self-reports was very similar pointing to a trait explanation of positivity bias (or at least domain-general character of the bias). However, the evidence provided by Grzegorczyk (1978) or Rynko and Palczyńska (2018) disagrees with this statement and favours domain- and context-related explanations. The domain-specific account is also supported by the results obtained by Anderson, Srivastava, Beer, Spataro & Chapman (2006) and a comprehensive review by Beer and Harris (2019). Most probably, the interactionist account of Ziegler (2015) claiming that both context and traits play role would eventually prevail here[36] as, as it was claimed by Paulhus and Trapnell (2008), the tendency to self-enhance is always present (though there is possibility that its magnitude varies across-individuals), but it is triggered by situational factors and the exact content of the self-presentation is context-dependent.

3.5.2 Positivity bias in educational context

Empirical analyses of this work are based on the PISA 2012 database, hence it is warranted to bring more focus on positivity bias in the educational/school-related context before presenting the main results of the study. The most typical paradigms in the school context positivity bias are comparing students self-reports to a) teacher ratings or b) objective test scores (Brown, Andrade & Chen, 2015). Positive illusions in school context were associated with poorer social skills, more problem behaviours and lower academic achievement (Gresham, Lane, MacMillan, Bocian & Ward, 2000). Other researchers linked self-enhancement bias with narcissism, poorer social skills and lower academic achievement as measured by lower grade point average (GPA) (Kwan, John, Robins & Kuang, 2008). Kim and colleagues (2010) related self-enhancement to lower life satisfaction and GPA among inaccurate self-assessment group (Study 4 & 5). Forsterling and Morgenstern (2002) obtained lower scores of motivation to achieve among inaccurate self-assessors. Gramzow and co-workers also corroborated the negative relation between self-enhancement and academic achievements, e.g. GPA (Gramzow, Elliot, Asher & McGregor, 2003). However, there are also results pointing to positive consequences of overoptimistic self-reports, e.g. lower drop out from school and higher teacher ratings of performance (Bonneville-Roussy, Bouffard & Vezeau, 2017).

[36] As it is very often in social sciences, cf. “nature or nurture”, “state vs. trait” or “conscious vs. unconscious” debates from various research contexts.

In general, it is suggested that students of lower academic skills have also lower abilities to self-assess themselves adequately and that boys tend to have more self-enhanced views than girls. Moreover, the self-reports validity in case of most students tend to improve in the course of education (Brown et al., 2015), probably as a function of feedback provided by teachers and parents (Chung, Schriber & Robins, 2016). The proportion of overly positive self-reports is assessed to be around 30%, with around 15% of students showing self-effacement (Bonneville-Roussy et al., 2017). Interestingly, Minkov (2008) showed that on the country-level self-enhancement tendencies may lead to lower academic achievement as indexed by results from PISA and TIMSS studies. He also suggested few possible explanations for such result: a) negative relation between self-enhancement and self-improvement motives, resulting in self-complacency and lack of motivation to improve one's skills (see also Heine, 2003), b) ignoring negative information more in more self-enhancing countries, c) setting lower objective achievement standards that lead to inflated self-reports in the cross-country perspective and d) counter-productive school practices promoting superficial criteria of achievement, e.g. norm-referenced superiority (being the best in class) and inducing mostly external motivation, e.g. learning to achieve good grades or to be "the best in my class" (Minkov, 2008; see also Watkins, McInerney, Akande & Lee, 2003). So far the causal order of this postulated relation is not known.

3.5.3 Adjustment of positivity bias

An important debate in the field concerns outcomes and (alleged) adjustment benefits of self-enhancement. In general, it is believed that self-enhancement brings positive social consequences, at least at short distance. Adaptive benefits of self-enhancement were advocated by Taylor and Brown (1988) on the basis of negative correlations between SDR scales and health problems, anxiety, negative affect and other similar measures. Other evidence is not so favourable for this hypothesis as e.g. Colvin and Block (1994), Colvin, Block and Funder (1995), Martocchio and Judge (1997), Paulhus (1998), Peterson and colleagues (2003) suggested that self-deception and self-enhancement were maladaptive. Dale and Weinberg (1990) contended that self-enhancement could lead to unrealistic ambitions and consequential failure, frustration and burnout. Similarly, self-enhancement may lead to disengagement, as the pain of failed high (but unrealistic) expectations is too big (Robins & Beer, 2001). Gorlin and Otto (2017) even saw self-deception as a major threat to psychological integrity and saw it as one of the major challenges before psychopathological research, proposing even organising therapies dedicated for self-deceivers.

An interesting innovation was proposed by von Hippel and Trivers (2011), and later confirmed empirically by Smith, Trivers and von Hippel (2017), who claimed that self-deception (positivity bias) do not serve for *intra*personal benefits but it is mainly oriented toward *inter*personal gains, such as winning social respect. The research by Anderson and colleagues (2012) showed that individuals who display confidence, even if it is unwarranted by their knowledge and abilities, gain higher social status and are perceived as more competent than they really are[37]. Mijović-Prelec and Prelec (2010) also noticed the relation between positivity bias and confidence boost. Also the research of Smith and collaborators (2017) showed that self-deception's benefit entails in being more persuasive, as self-deceived individual is more convincing and sways people more easily. In this way self-deception also masks manipulation and lowers the chances that an individual will be revealed as incompetent or manipulating. The negative consequences of unsuccessful self-presentations are well-known in the literature and have been described e.g. by Beer and colleagues (2013). The research up to date also

[37] Anderson and collaborators use the term "overconfidence" to describe state of unwarranted confidence in one's abilities. This sense of this term is a bit different from its use in survey methodology field where it is used to denote metacognitive bias in assessing confidence of one's predictions (Stankov & Crawford, 1997). However, both uses are related to each other as both of them denote a gap between perceived and actual abilities.

points that self-enhancers are more liked by others, but it is not universally confirmed by all research results. An important thing is to differentiate between actual and perceived self-enhancement, as the latter is most often not acceptable, namely when others realise that an individual is self-enhancing (thus, deceiving them) they tend to frown upon at such insincere self-presentation (Dufner et al., 2019; Sedikides, Hoorens & Dufner, 2015). This is probably why recent meta-analyses suggested that self-enhancement can have positive social consequences only in short-term, with negative consequences in long-term- simply in the long run there is greater probability that self-enhancing, overconfident individuals be uncloaked and frown upon for their self-enhancing attempts (Chance et al., 2011; Dufner et al., 2019). Both meta-analyses suggested that long-term benefits of positivity bias may be sustained, if the bias is yielded only in the private context. However, results brought by Beer and Harris (2001) showed negative long-term consequences of self-enhancement also in the intrapersonal domain, leaving the evidence gathered to date without firm conclusions (Beer & Harris, 2019). Also in the case of self-enhancement in the school context there are no firm conclusions about the adjustment of overly optimistic self-reports of abilities as the evidence is mixed and some studies point to positive outcomes like increased educational attainment or school retention, whereas other point to negative consequences (Brown et al., 2015).

In the line of other-deception research an especially interesting account was presented by Lamba and Nityananda (2014) as they reviewed possible consequences of self-deception for financial and educational institutions. The researchers claim that gains from other-deception are potentially disastrous for societies as false air of competence surrounding self-enhancers may lead to their unwarranted promotion for key positions in important institutions like banks or armies leaving them in the risk of being run by incompetent but overconfident, others-manipulating individuals. Lamba and Nityananda advocated for taking this into consideration, e.g. in hiring decisions or grading in educational institutions (2014). They also call for taking this matter under consideration in scientific research and present initial results that ranking rather than grading might prove to be less prone to overly positive self-perceptions.

It is worthy to note that in many research attempts measures used as indications of self-enhancement consequences might be themselves distorted by self-enhancement (self-reports of health or self-esteem) or its social consequences (peer-ratings, social position) (Humberg et al., 2019). Also the magnitude of bias may be a key factor in judging its adjustment benefits. Mijović-Prelec and Prelec (2010) as well as Papps and O'Carroll (1998) were among the researchers who turned attention to the magnitude of positivity bias and claimed that moderate levels of self-deception may be beneficial or at least harmless. Other important factor may be the precise strategy used to achieve an enhancing effect. Hepper, Gramzow and Sedikides (2010) distinguished almost 60 different, both cognitive and behavioural, strategies, grouped empirically in four main factors. As showed by their research it is warranted that different strategies yield different, positive or negative, adjustment outcomes.

Additionally, more and more newly appearing research attempts point that the relation between abilities and adjustment is often mixed or curvilinear. Firstly, it may be domain-dependent, as, for example, in the case of emotion recognition that may be beneficial for job performance, as it is handy in negotiation processes, but may be detrimental for relationship maintenance due to increase stress in negatively valenced situations (Simpson et al., 2011). Similarly cognitive abilities may facilitate work performance but may hinder gaining leadership in group, as often groups refrain from having overly smart leaders (Antonakis, House & Simonton, 2017). Thus, self-enhancing consequences may be different depending on the exact context, domain and shape of relationship (He & Cote, 2019). As posited by Thompson (1999): in one context self-enhancement may be beneficial, e.g. by reducing stress reaction (cf. Why & Huang, 2011) or by motivating individual to increase effort (self-efficacy

explanation, Bandura, 1997), while in the other it may be maladaptive, e.g. due to threat underestimation and subsequent risky behaviours or due to causing helplessness (cf. Donovan, Leavitt & Walsh, 1990).

3.6 Chapter summary

The first part of the above chapter concentrates on presenting evidence on positivity bias and related topics stemming from different research fields and methodologies than the "mainstream" of socially desirable responding research presented in Chapter 2. The integrated evidence from e.g. observational research of Erving Goffman and survey research of Elisabeth Noelle-Neumann largely corroborated findings from the main body of research.

Some interesting results were brought by Peter Blau who turned attention to higher frequency of positive in comparison to negative feedback in social relations which may be one of the reasons of the prevalence of positivity bias. The findings presented by Noelle-Neumann (1974) linked SDR to conformity and evidenced that people are constantly gauging what is socially desirable and socially approved in their particular groups. Therefore, what is perceived as socially desirable can change from group to group, e.g. from school to school. Very thought-provoking results were obtained by Bishop (1986) who showed that people can claim opinions and knowledge on non-existing, entirely fictitious issues. This urges the question about the boundaries of self-presentation and self-deception but also provokes queries about the mechanisms that lead to such effects.

Interesting information on the nature of positivity bias was brought by psychological research on self-consciousness, self-knowledge and self-motives. Most importantly, it was showed that achieving accurate self-knowledge, indispensable for forming realistic self-judgments, is an effortful and difficult process that requires sufficient control and cognitive resources to be brought off successfully. Moreover, at least for some participants, gathering accurate self-knowledge can be even aversive and unpleasant. Forming veridical self-judgments also requires overcoming automatic reactions, that promote (overly) positive self-view and tune attention to positive feedback and desirable outcomes (see e.g. Gesiarz et al., 2019).

The above chapter also presents a thorough review of self-enhancement and both motivational and non-motivational accounts of this phenomenon. The former concentrates on the roles of self-enhancement as a boost to self-esteem, protection from threatening stimuli or thoughts, energising factor, overcoming fear and anxiety and enhancement social status. The latter sees self-enhancement as a by-product of cognitive processes, e.g. as a result of limited cognitive resources or memory fallacies. Interestingly, evidence from neurocognitive research shows that self-enhancement as a reaction to self-esteem threat and as a result of cognitive load are based on different neuronal basis. Furthermore, this branch of research claims that self-enhancement is one of the key motives that guide everyday behaviour and cognition. Hence, positivity bias can be expected in almost every measurement occasion.

Investigation for correlates of positivity bias leads to a mixed picture, pointing also that pairwise correlations are unlikely to shed much light on the relations between positivity bias and psychological and socio-demographic characteristics. As evidenced by research (e.g. by Rynko & Palczyńska, 2018) such relations should be modelled by interactional designs as suggested by theoretical models of Ziegler (2011) and Krosnick (1991; 1999). Nevertheless, positivity bias was related to a wide array of socio-psychological traits, e.g. self-monitoring, dark personality, overconfidence, low educational attainment, but also to measurement occasion characteristics: domain desirability, verifiability and individual importance (centrality) among others. In the educational context specifically, lower

educational achievements, lower school motivation, more behavioural and social problems were correlated with more bias. Moreover, boys typically yield more enhanced responses than girls.

As it seems from the research amounted to date there is a tremendous need to focus more on measures used to index positivity bias and establish its correlates (Beer & Harris, 2019; He & Cote, 2019). Moreover, the characteristics of measurement context (e.g. accountability, private vs. public, cognitive load) and items (desirability, verifiability, etc.) should be experimentally manipulated and linked with other self-report measures and also other, not self-descriptory tasks in order to advance knowledge on positivity bias. Specifically, an approach to overcome the shared method variance concern are needed as often trait and method variance are confounded as measures of positivity bias and measures of outcome come from the same source (most often the participant himself). Additionally, more longitudinal studies with diverse indicators of adjustment, preferably using objective measures, including psychophysiological or neurocognitive data, is needed. In their comprehensive review article Beer and Harris (2019) sum up that these steps are needed "to understand the nuances of when, where, and for whom self-insight failures are adaptive". It can be added: "when, where and for whom self-insight failures happens and what effects they have on self-report validity".

Some clarifying evidence was recently provided by Garrett, González-Garzón, Foulkes, Levita and Sharot (2018) who claimed that positivity bias disappears in threatening conditions. Hence, the whole evidence accumulated in laboratories and surveys, all collected in non-threatening, non-stressful conditions may be only one facet of the self-enhancement relation to adjustment. The other facet may be reversing the preference for positive stimuli under a perceived threat, leading to disappearance of positivity bias. Such change in preferences may be executed by a shift in attention and was corroborated in the EEG research of Carretié and collaborators (2004) who showed that processing negative stimuli is quenched on early stages of processing if they are not identified as a real threat, precisely as suggested by the research of Garrett and colleagues (2018), Kappes and collaborators (2020) or Rigney (2019)[38]. However, this pattern can be reversed when a situation would be identified as truly threatening, as suggested by new evidence (Garrett et al., 2018). Interesting results in this context were brought by Cai, Wu, Shi, Gu and Sedikides (2016) who showed that higher processing of negative stimuli, and consequential decrease of self-enhancing responses, is possible. Such processing was marked by larger N170 event-related potential (ERP) as a response to negative information in less self-enhancing participants.

Hence, it can be concluded that positivity bias can be triggered at least in two ways: either by a self-relevant motive (Sedikides, 1993) or by impaired cognitive processing due to e.g. cognitive load or time pressure (Kappes & Sharot, 2019). The two processes are distinguished on the neuronal level but it is not known whether they also yield different behavioural effects (Beer & Harris, 2019; Flagan & Beer, 2013). More is known about the motivated positivity bias which is an automatic, non-deliberate and adaptive phenomenon of human cognition that serves for both intra- as well as interpersonal benefits, e.g. stress reduction (Hernandez et al., 2015), motivation and self-efficacy boost (Bandura, 1997), as well as winning social support and status (Anderson et al., 2012; Smith et al., 2017). However, unmitigated or uncorrected positivity bias can lead to negative consequences, both intrapersonal (e.g. lowered self-esteem; Robins & Beer, 2001) and interpersonal (e.g. lowered social liking; Paulhus, 1998). This view of positivity bias accounts perfectly for the mixed evidence on its adjustment and consequences- simply put, the bias brings positive outcomes when it offers net benefits and when the net balance is no longer favourable it should be, and most often is, corrected (Johnson & Fowler, 2011).

[38] It is worthy to note that the notion of rejecting all self-threatening information was already suggested by Hilgard (1949).

Therefore, positivity bias brings positive consequences for most of the people in most of the situations. However, survey methodologists are not among the fortunate and positivity bias always brings them (potentially) negative outcomes. What are the consequences of self-enhanced responses in self-report data and what methods were proposed to curb them is presented in the subsequent chapter.

Chapter 4-SELF-REPORT METHODOLOGY AND POSITIVITY BIAS: IDEAS, PROBLEMS, REMEDIES

4.1 The idea of self-report of abilities

The idea of self-report is beautiful and simple- why bother with complicated measures if one can just ask participants about virtually anything that warrants scientific assessment? Moreover, what better way to measure so many hard to observe phenomena? Lastly, who could possibly know their personality, attitudes or math abilities better than the respondents themselves? Unfortunately, this rose-coloured situation is only half true- self-reports offer tremendous benefits but are also prone to many threats to validity, response biases among them.

Self-report scales are constructed on the basis of assumption that participants are able to validly report on certain existing characteristics (traits) with the use of a measurement instrument provided, most often in the form of a rating scale. However, Mills and Hogan (1978) called this basic assumption a "Platonist interpretation of item responses" claiming that already the assumption of the traits existence is an optimistic idea[39], not to speak about the possibilities that participants can veraciously describe themselves. Mills and Hogan (1978) also pointed out that measurement occasion is a social situation for respondents, meaning that they will employ role-taking (role-playing) behaviours just as they would do on any other occasion. This implies that self-report measurement is under threat of all the conscious and unconscious tendencies predicted by theories of Goffman, Blau, Hogan, Paulhus or Sedikides, as presented in the above chapters. Shortly, respondents strive to maximise their positive experiences and minimise their negative ones during any measurement occasion and will yield an overly-positive image of their characteristics due to inherent response biases, e.g. impression management, self-enhancement motivation or simply putting low effort in the responses provided. Problems with self-report measurement typically present themselves regardless the domain measured, be it reporting about frequency of physical exercises, dietary habits, one's personality or reading abilities (Brenner & DeLamater, 2016; McIntyre, Noels & Clement, 1997). However, despite many potential pitfalls the self-reports manage to yield quite high validity when compared with more "objective" measures. The subsequent chapter will also briefly present results of validity studies in the context of self-reports and will also provide a short overview of the methods conceived to control for positivity bias. In both cases the presentation is focused on self-administered self-reports as this is the focal point of the subsequent analyses.

4.2 The validity of self-report

In a recent study Naguib and co-workers (2019) conducted a large-sample web survey among medical doctors. On average, the respondents achieved only 57% of items answered correctly on a 9-item test, whereas the mean predicted score amounted to 84%, yielding that 92% of the surveyed doctors were overoptimistic about their skills. This effect could be blamed on the "ignorant, but overconfident experts" (Bradley, 1981) but other, more lay samples also provided some discouraging examples. For instance, Ennis (1965) surveyed participants about their reading activities and, probably to his great dismay, obtained results that even people that reported reading zero books in the last year assessed

[39] The researchers here referred to the notion of "traitedness", namely, how similar a person is across situations or what is the part of the behavioural or attitudinal variance explained by situational versus trait factors (Fleeson & Wilt, 2010; Sheldon, Ryan, Rawsthorne & Ilardi, 1997). The "traitedness" is also thought to vary cross-culturally (e.g. Minkov, 2008). See also Block (1961) on ego identity and role variability.

themselves as “moderate” or “heavy” readers (18% of the “zero” group). Another 38% of the group that reported reading 1-4 books per year rated themselves as “moderate” or “heavy” readers[40].

Probably a parade of such results pushed DeNisi and Shaw (1977) to conclude that self-reports of abilities are impractical as they did not correlate with ability tests. Although the data they gathered did not support such a harsh claim as self-reports correlated with tests, though not every scale that should did and some correlations were low in magnitude. Thornton (1980) ascribed these lack of predictive validity to an “inflation bias” and admitted that the evidence of self-report predictive and discriminative validity is mixed. However, it is worth to note, that DeNisi and Shaw (1977) based their study on a sample of students in a low-stakes assessment, whereas other studies that recruited real job applicants or real job employees (high-stakes context) have shown that self-reports of ability, knowledge and skills are a valid predictor in personnel recruitment (e.g. Levine, Flory & Ash, 1977), yielding high correlation between self-reports and objective ability tests and other measures, e.g. supervisors- or peer-ratings. Thus, context and instrumentality, as well as stakes, could have crucial impact on the usefulness of self-report measures. Furnham (1986, 1990) also claimed that measurement tools differ in their susceptibility to faking due to their design. He believed that scales with high face validity and these measuring constructs well-known or well-understood by the general public being in more risk of being faked. This claim has a naturally logical appeal and has been also tested empirically (Furnham, 1990; Furnham & Henderson, 1982; Velicer & Weiner, 1975), however it is unknown how these scale characteristics would interact with positivity biases other than faking.

Such discrepant results urged researchers to perform integrative studies and meta-analyses in order to assess self-reports’ utility as well as to investigate potential moderators of their validity. The meta-analyses in general have confirmed that self-reports of abilities are valid measures of constructs of interest, as most of the meta-studies have yielded that self-descriptions correlate with more objective measures, mainly cognitive tests (e.g. IQ tests, math proficiency tests, etc.). Mabe and West (1982) reported a self-report-objective measurement correlation to be low to moderate (r=0.29 on average) with high variability (SD=0.25). Some scales yielded higher correlation with their criterion (even up to r=.70) and some failed to present any criterion-related validity at all (correlations around 0). Very similar results were presented by Hansford and Hattie (1982) who obtained average correlation of 0.21 and again high variability of correlation coefficients of (SD=0.23)[41]. Similar results were also obtained by Kruger and Dunning (1999) who claimed that in most research endeavours self-reports yielded only moderate correlations with actual performance (rs from 0.05 to 0.50). Research on correlation between SAT scores (objective measurement) and self-rated ability yielded only a 0.33 correlation, GPA (grade point average) and self-rated ability- only 0.22 (Robins & Beer, 2001). Kim et al. (2010) reported higher correlations between these measures (r~0.60-0.80), but with a much shorter test, although also based on SAT items. Joseph and Newman (2010) presented a meta-analysis of relations between self-report and various objective measures of emotional intelligence and yielded a correlation of only 0.20. Also research on the utility of the single item IQ measures yielded their low utility as

[40] These results were also wittily depicted in the pop culture, e.g. in the Woody Allen’s movie “Annie Hall”:
Alvy's Psychiatrist: [Alvy and Annie are seeing their therapists at the same time on a split screen] How often do you sleep together?
Annie's Psychiatrist: Do you have sex often?
Alvy Singer: [lamenting] Hardly ever. Maybe three times a week.
Annie Hall: [annoyed] Constantly. I'd say three times a week.
[excerpt from the screenplay of the movie as in https://www.rottentomatoes.com/m/annie_hall/quotes/]
See more on this kind of effects of unclear terms (“vague quantifiers”) but also inter-group comparisons in the works of Nora Schaeffer, e.g. Schaeffer & Charng (1991).

[41] Educational settings’ data was analysed in that study.

intelligence proxies as the correlations between the scores were low (*r*~0.20-0.25), despite the strains to improve the validity by manipulating item design (Paulhus, Lysy & Yik, 1998). However, specifically in the domain of math self-assessment the correlations between self-report scales and objective tests can amount up to 0.50 (Ackerman, Beier & Bowen, 2002), being a high value in the light of a recent meta-synthesis[42] (Zell & Krizan, 2014).

Other researchers claimed that it is also important to accept that every self-assessment measure would be always imperfect, in the sense it would be always discrepant with other, more objective measures (John & Robins, 1994; Kim et al., 2010). However, when self-assessment and objective test are on the same scale (e.g. number of points) the correlation can be quite high (*r*~0.60-0.80) and around 80% of the sample will yield correct estimates of one's score (Kim, et al., 2010). The evidence gathered shows that respondents have a limited insight into their own abilities but that there is also a significant importance of factors that moderate the criterion-related validity of self-reports (Zell & Krizan, 2014).

4.2.1 Moderators of self-reports validity

Mabe and West (1982) differentiated moderators into: a) person-level characteristics and b) measurement conditions, the former entailing individual differences, such as age, gender, intelligence or personality, whereas the latter regarding item design, domain characteristics or measurement context.

Person characteristics

The first and obvious candidates for a moderator of self-reports criterion-related validity are response biases with a special focus on the positivity bias. In fact, there is evidence that successful control of the spurious error variance leads to validity improvements (e.g. Anderson et al., 1984; Leite & Cooper, 2010). However, pinpointing such variance is not an easy thing and oftentimes even (theoretically) very elaborated tools fail to account for it and do not moderate the criterion-related validity (DeNisi & Shaw, 1977; Li & Bagger, 2006; Huang, 2013; McCrae & Costa, 1983).

Another often researched moderator is age. In general, there is evidence that even young children (ages 8-12) yield overly positive self-descriptions (Thomaes, Brummelman & Sedikides, 2017) but the older the sample the higher the validity of self-reports (Paulhus et al., 1998). This pattern is confirmed by many research projects and has a solid grounding in the psychological theory of cognitive development (Harter, 2012).

Gender also was scrutinised as a potential moderator, but the research suggests that it can be a validity moderator only in certain domains, related to gender stereotypes and social norms in a given society (Borgonovi & Pokropek, 2019; Marsh & Yeung, 1998).

Surprisingly, there are not many evidence-based personality moderators of self-report validity. One of them is internal locus of control (Mabe & West, 1982), which is attributed to superior utilisation of personally-relevant information and also to greater motivation to actively seek such information in comparison to respondents with external locus of control (Phares, Ritchie & Davies, 1968; Seeman, 1963). In general, the research on personality moderators is scarce (Ackerman et al., 2002; Mabe & West, 1982).

Personal cognitive abilities are also an important moderator of self-report validity: more intelligent and more educated participants yield responses of higher criterion-related validity (Mabe & West, 1982; Schlosser et al., 2013). Role of metacognitive ability as well as experience in self-assessment and

[42] Meta-analysis of meta-analyses.

survey participation was also researched. It was suggested that there is no gain from experience as participants failed to yield higher self-report validity in a post-test in comparison to first measurements (Ferraro, 2010; Hacker, Bol, Horgan, and Rakow, 2000; Schlosser et al., 2013) but the research presented by Miller and Geraci (2011) and Ryvkin, Krajc & Ortmann (2012) provided evidence that corrective feedback and metacognitive training can result in larger validity of self-reports. Experience in survey participation was positively related to data quality (Gadzella, Cochran, Parham & Fournet, 1976; Gummer et al., 2018; Heilenman, 1990). These results indicate that accurate self-assessment can be taught, at least to some extent (Butler & Winne, 1995; Miller & Geraci, 2011).

Interest and motivation were linked with higher data quality (Campbell & Lavallee, 1993). These characteristics can be both specific ("I am interested/motivated in this particular topic") and general ("I am interested/motivated to take part in surveys/acquire knowledge of myself/help science in general") as suggested by Paulhus et al. (1998).

Measurement characteristics

Specific *versus* broad items give estimates of higher validity (Ackerman et al., 2002; Mabe & West, 1982; Zell & Krizan, 2014), e.g. item "I am good at math" is predicted to yield ratings of lower validity than item "I can calculate the area of a triangle". This relation is probably driven by how the self-efficacy is created which is inherently a task-specific trait (rather than task-general or domain-general) as it is related to task-specific exposure and experiences (Borgonovi & Pokropek, 2019).

Difficulty of the task also may play a role, with hard tasks being more often over-rated and easy tasks-under-rated (Juslin, Winman & Olsson, 2000). The results of the meta-synthesis provided by Zell and Krizan (2014) suggested that self-reports on simpler, less complex tasks yield higher validity than reporting on more complex tasks.

Task/domain familiarity is another measurement characteristic positively related to criterion-related validity (Mabe & West, 1982; Zell & Krizan, 2014). It is postulated that familiarity enables learning process to take place, which eliminates at least part of spurious variance (Butler & Winne, 1995).

Intuitively more objective, verifiable domains are self-reported with more accuracy than subjective, unverifiable ones (Mabe & West, 1982; Zell & Krizan, 2014). Thus, it can be expected that math ability will be more accurately reported than e.g. emotional intelligence (Paulhus et al., 1998).

Trait desirability is related negatively to the validity of self-reports- surveying about subjectively important or socially desirable traits is a typical situation where self-enhancement motivation can lead to biased scores (Paulhus et al., 1998)

Stakes of the assessment are also related to the validity of research. Mills and Hogan (1978) predicted that in low-stakes situations validity should be higher as incentives to faking or self-enhancement are low in these situations (see also Niessen et al., 2017). On the other hand, it is postulated that high-stakes assessments may be less biased as they induce additional motivation (e.g. Eklöf, 2010; Eklöf & Nyroos, 2013).

Both anonymity and accountability are related in general to the validity of the self-reports (Mabe & West, 1982) as it was already commented in subchapter 3.5.1 on the role of accountability in self-enhancement reduction.

Certainly this subfield suffers from lack of research, especially scarce are experimental studies (Mabe & West, 1982; Zell & Krizan, 2014). It seems that better knowledge on moderators of the validity has been acquired in case of cognitive tests than self-reports (see e.g. Borgonovi & Biecek, 2016).

4.3 Problem with "objective" criterion

Obviously, criterion-related validity studies have their own, specific problems. The first group of them is of purely psychometric nature. Put simply- psychometric qualities of the tasks used both as self-reports and objective measures are another important moderator of criterion-related validity (Reynolds & Suzuki, 2013). If unreliable or lacking in construct validity measures are used in research then the correlation coefficients are probably limited or even distorted (Ackerman, 1996; Paulhus et al., 1998). Another issue of similar nature is the problem of method *versus* trait variance: many self-reports are correlated higher with other self-assessments than with other, e.g. behavioural, measures of akin constructs indicating a large role of response biases[43] (Campbell & Fiske, 1959; Joseph & Newman, 2010; Podsakoff et al., 2003).

The second group of problems is of more philosophical nature. The criterion-related validity studies entail using the best possible criterion in order to obtain values close to the "true score". However, not in every measurement it is clear what this "true score" or "true value" should be. According to many researchers the distinction is clear-cut in case of socio-demographic variables (e.g. gender, age, profession) or questions about objectively verifiable facts (e.g. number of medical visits in the past year, income from all sources in the past month, etc.)- there is a certain value or score that is true and any distortion from it would be considered an error. However, in case of many measures it is hard to establish what the "true score" should (and could) be. This problem is especially prominent in measuring constructs like opinions, attitudes and judgements (Groves, 1989; Jabkowski, 2015; Sztabiński, 2011). After all, it is possible that there is no "true" value outside of the measurement situation as in many cases the attitude in question is constructed simultaneously with the measurement process (Jabkowski, 2015).

This situation is best reflected in the difference between the "true score" notion in psychometric *versus* in statistical sense (Jabkowski, 2015). In statistical definition the true score is simply the value of an index in the population, whereas in psychometric sense the true score is latent or even non-existent (Groves, 1989). Hence, under the psychometric definition any value without error could serve as a measure of true score, whereas in the statistical definition the true score is simply the value of an index in population. Despite the examples provided by the researchers it is still unclear to which category measurement of certain constructs should be classified. In example, are cognitive abilities closer to the elusive nature of political attitudes or are they rather affined to the firm factuality of socio-demographic variables?

Setting this notion aside it is warranted to conclude that in many cases theoretical validity can be only estimated not determined. However, it is crucial for any validity study to provide "best, most informative, least ambiguous evidence that resources allow" (Landy, 1986). In a criterion-related validity study the main focus is dedicated to finding the best possible criterion, otherwise the notion of validity as a correlation between indices is seriously limited. The PISA math ability test fulfils these high requirements for a "representative and pre-eminent" (Landy, 1986) or an "exemplificatory" index (Groves, 1989; Jabkowski, 2015) of math ability due to its high psychometric characteristics and careful preparation of the tests' content (OECD, 2014a; 2014b). Moreover, it offers measurement conducted on a representative, large sample in standardised, controlled conditions and with strict rules of data processing after measurement. All these facts attest that the PISA math test scores are as good a criterion of math abilities as it is feasibly thinkable in case of social sciences research[44].

[43] And other measurement errors related to the form, but not to the substance.

[44] Of course, the PISA test has also many drawbacks, among which being a low-stakes test is one of the most important ones (Borgonovi & Biecek, 2016; Eklof, 2010; Rutkowski & Wild, 2015). Another possible source of

Hence, the correlation between the test score and the math familiarity (self-report) score will be used as the main indicator of criterion-related validity of the math familiarity scale in order to assess the effect of OCT on this relation. This is an example of a concurrent validity study as both measures were taken at the same time point. This is also a validation of "X as a sign of Y" type (prediction validity), but also "X requires the same attribute as Y" (construct validity) (Landy, 1986) and OCT score is used to assess the magnitude of response bias in the self-report.

4.4 Commonness of overly positive self-reports

There is evidence that inflated ratings are very common indeed (Shi, Sedikides, Cai, Liu & Yang, 2017). Brandt (1958) concluded that only 50% of self-estimates were adequate, others were mostly too high. Very similar evidence was provided by John and Robins (1994) who found that only 50% of participants were accurate in self-evaluation, 35% overestimated their performance, whereas 15% self-effaced (underestimated their performance). Preliminary research by Koniewski et al. (2019) revealed slightly lower numbers with only 4% of self-effacers and 16% of self-enhancers on both tasks they have used but with greater numbers for one-task self-effacement/enhancement (21% and 39% respectively). Similar numbers in the school-context studies were reported by Robins and Beer (2001) where 31% of the sample overestimated their performance and only 9% underestimated and Bonneville-Roussy and colleagues (2017) were almost 30% of the sample yielded inflated self-reports but also around 15% yielded overly pessimistic ratings. Almost exactly the same proportion of 30% of overclaimers was obtained by Randall and Fernandes (1991). Similar proportions of participants distorting their self-ratings were yielded by faking studies where the percentage of fakers oscillates usually between 15 and 30% (Holden & Book, 2012). However, the average size of the positivity bias is usually not large (Robins & Paulhus, 2001), most of the people yield only mild to moderate bias (Alicke & Govorun, 2005; Dufner et al., 2019; Robins & Beer, 2001) and only a fraction yields large bias (Taylor & Armor, 1996). In the meta-analysis performed by Viswesvaran and Ones (1999) the average effect sizes pointed out that most of the fakers had scores elevated by a half or one standard deviation, depending on the scale. Hence, it is important to differentiate between a "bias" and a derailment of cognition as the two are definitely not the same, the latter indicative more of clinical problems than self-enhancement.

Most of the faking research undertaken so far shows that faking can lead to distortion of personality assessment rising up to one standard deviation (Birkeland et al., 2006; Dwight & Donovan, 2003; Hooper, 2007; Viswesvaran & Ones, 1999). Moreover, various research endeavours suggest that proportion of participants engaging in faking may vary from around 15% to as high as 50% (Ziegler et al., 2011). Although the influence of faking on self-descriptory data in other research areas is not that meticulously researched, there is sufficient evidence to conclude that they are not free from its detrimental influence. For example, in the study of Taylor and Brown (1988) none of the surveyed students expected to score below the mean on the final exams, whereas in a study of Krueger (1998) almost half of participants said they would be capable of solving a complex mathematical problem that only 20% of population would solve. Similarly, 88% of the surveyed American sample claimed better than average driving skills (Svenson, 1981) and 94% of surveyed academic teachers claimed above-average teaching skills (Cross, 1977)[45].

error in the PISA databases is related to cheating, faking and other errors, including those committed probably in the data processing phase (see Blasius & Thiessen, 2012; 2015), however, this factor is possibly present in many other datasets as well.

[45] Those research results are classical examples of the so-called above-average or better-than-average effect and are most probably effects of self-cognition biases (self-enhancement), leading to ascribing oneself more positive traits and skills than it is in reality (Thornton, 1980; Wojciszke, 2011).

Participants not only assess themselves more favourably but also claim to have traits or to perform actions that in reality they do not perform, nor have. In a population-representative study conducted in Poland around 75% of the English teachers surveyed claimed that they use information-communication technologies (ICT, e.g. computers) during their lessons, whereas observations of their classes yielded that only 25% of the teachers surveyed actually used them (Muszyński, Campfield & Szpotowicz, 2015). Moreover, in the Polish edition of the PIAAC study around 20% of participants declaring frequent use of computers in a self-descriptory survey then failed a very simple computer proficiency test (Burski et al., 2013). In the PISA 2012 study (OECD, 2014a) even 15% of pupils claimed that they are very familiar and understand well non-existent mathematical terms (e.g. "indicative fraction") and more than 30% of them claim that they have heard about those faked terms. What is important, is that the results of those self-descriptory data on mathematical abilities have rather low predictive validity on mathematical test results. Only accounting for scores on non-existent items led to rise of the validity of the self-descriptory data (Pokropek, 2014). Those results show that virtually any use of self-descriptory scales is in a threat of faking and that basing research conclusions on uncorrected data is in a risk of drawing wrong conclusions.

4.5 Review of SDR control methods

4.5.1 Problems and limitations of the actual SDR methods research

Many SDR "reduction" methods have been developed, although despite a large amount of studies as yet none of the proposed palliatives gained universal recognition as a commonly-accepted standard to deal with positivity biases (faking, SDR, etc.) (Krumpal, 2013; Ziegler et al., 2012). Moreover, there is still a significant gap in efficiency comparative studies where various methods are pitted against each other. Validity studies are also scarce, especially criterion-related validity research. Another problem is that the majority of the studies devoted to this topic where conducted in only one specific setting: faking of personality assessments, mainly in practical contexts. Other topic highly covered in the literature is controlling for SDR in sensitive and intrusive questions. The remaining topics are underrepresented in the literature which results in a significant tilt of suitability of the existing methods to the two most commonly researched situations. However, as has been shown in subchapter 4.3, positivity biases are not limited to these contexts. The concentration on practical utility in a narrow number of contexts resulted in a relative lack of theoretical advancement in the field which impedes further research (Mueller-Hanson, Heggestad & Thornton, 2006; Ziegler, 2011). Nonetheless, these problems have been identified and new research endeavours aim at "putting the horse back in front of the cart" (Heggestad, 2012), namely trying to link practical method utility with substantial understanding of the basis of their effectivity.

The subsequent review will try to present most important methods devised to control positivity bias, along with an attempt to gauge their adequacy for large-scale assessments and self-reports of abilities situations. The review concentrates on the methods possible to use in the self-administrative mode. Methods specific for situations where interviewers are involved are beyond the scope of this work. Exemplary information on this research can be found elsewhere (e.g. Bredl, Storfinger & Menold, 2011; Bredl, Winker & Kotschau, 2012; Kemper & Menold, 2014; Krumpal, 2013; Menold & Kemper, 2014; Nederhof, 1985).

4.5.2 Classifications of SDR control methods

Many classifications of positivity bias control methods have been proposed and almost every one of them entails division on methods intending to prevent the bias and methods aimed to reduce consequences of the bias that already took place. The preventive methods need to be applied before

the data is collected so they are often referred to as *ex-ante* or proactive methods, whereas the other group is applied after the data is collected, hence they are also known as *post-hoc*, reactive or remedy methods (Adair, 2014; Hipsz, 2014; Zheng, 2015; cf. Nederhof, 1985; Paulhus, 1991).

Other divisions proposed also take account of the level on which the method operates (person, item, scale, etc.; Dilchert & Ones, 2012; Kuncel, Borneman & Kiger, 2012) and the degree to which methods constitute part or by-product of the measurement of interest (so-called internal methods) or require using additional indicators (bias markers) with little or none substantial meaning (external methods) (Reeder & Ryan, 2012).

Adair (2014) proposed a classification based not only on practical aspects of the methods but also on the mechanisms they are aimed at. This classification integrated applied research with some earlier conceptions of faking conceived by McFarland and Ryan (2000; 2006). Adair divided preventive methods to these aimed at curbing participants' intentions to respond desirably (cf. demand reduction techniques in Paulhus, 1991) and these restricting their ability to do so (2014). The works of Adair (2014) and McFarland and Ryan (2000; 2006) were both dedicated for faking, but their extension to other positivity biases is proposed here. Obviously, these models are only in the initial phases of empirical testing (even regarding the faking framework only).

The model proposed by McFarland and Ryan (2000; 2006) derives from the Theory of Planned Behaviour (TPB; Ajzen 1991) and predicts that both situational and personal characteristics decide about the participants' intention to fake (respond desirably, self-enhance) that can result in actual faking (self-enhancing) behaviour if a given respondent possesses abilities to do so and the opportunities are favourable. The model is schematically shown in the figure below:

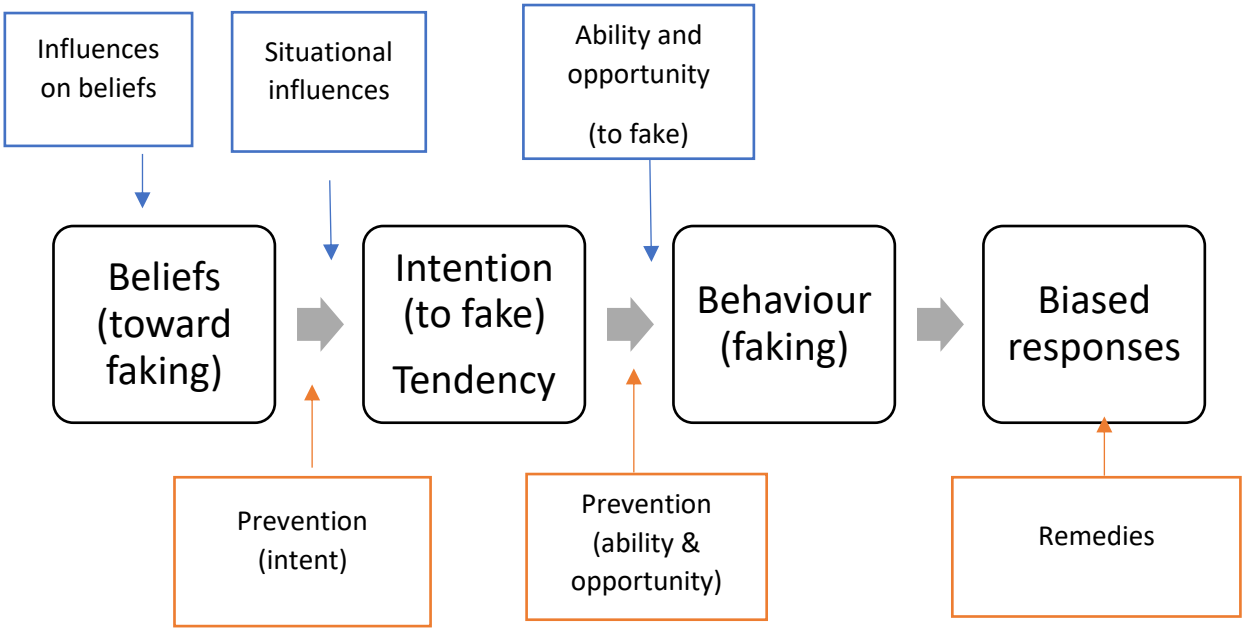

Figure 4. Model of positivity bias based on the model of faking proposed by McFarland and Ryan (2000). Methods and the stage on which they intervene according to Adair (2014).

McFarland and Ryan (2000; 2006) put a large emphasis on beliefs and intentions that precede faking behaviour in their theory. In their opinion such characteristics as values, morals or personality traits can influence the intentions to fake (e.g. "I will not fake as it is immoral"). On the other hand, even

having intentions to fake does not necessarily result in faking behaviour as it is assumed that participant may refrain from faking, e.g. as a result of adverse conditions (e.g. fear of detection) or insufficient abilities (e.g. lack of knowledge what self-image should be presented). Such a focus on beliefs and intentions is warranted in case of faking which is a conscious and purposive behaviour. However, it is also conceivable that similar processes and stages also take place when unconscious positivity biases take place. In example, participants' beliefs towards participating in surveys (e.g. survey/test anxiety) or even attitudes towards science in general can influence their measurement behaviour, even unknowingly to themselves. Similarly, in case of unintentional positive biases, "intentions of faking" could be changed to tendencies towards responding in an overly positive way. Research on self-enhancement confirms that such tendencies are ubiquitous but participants greatly differ in their realisation (e.g. Crocker & Park, 2004), hence warranting dependency of biased behaviour on preceding processes (intentions in case of faking and many other response sets, tendencies in case of response styles). Other parts of the model are also easily accommodated to the framework of general positivity biases (cf. Ziegler, 2011).

Returning to the classification proposed by Adair (2014): preventive methods need to suppress participants intention to fake or unable it through opportunity cancellation or ability reduction. On the other hand, remedial (identification) measures can be used only after the data is collected thus being able only to correct for bias effects, but not to preclude them. A detailed classification of positivity biases methods is presented in Table 1 below. Classification categories proposed by Adair (2014), Reeder and Ryan (2012) and Dilchert and Ones (2012) were merged and organised in order to formulate a comprehensive taxonomy.

It is noteworthy that this classification is only a proposition, though an important one. Identifying method's place in the classification enables inference about the mechanisms it is meant to affect. Of course, classification should be subjected to discussion and verification due to further empirical and theoretical developments. It seems that there is little place for controversy in the prevention *versus* remedy division, as well as in the intention *versus* ability or external or internal character of remedies. However, there is large overlap between the person-level and item-level SDR control methods. This overlap is rendered by introducing the "mixed-level" category which reflects that this classification is subjected to the highly contextualised use of these methods. In some studies randomised response technique may be treated as a person-level analysis as the questions submitted to randomisation procedure may constitute the whole or vast majority of measurement, whereas in other research attempts only a subset of items may be subjected to randomisation, meaning that this method was used as an item-level one (see e.g. Hipsz, 2014). Similar reservation involves also keying/weighting methods and differential item functioning (DIF)- they may be use as person-level indicators of desirable responding or they may remain only item-level indicators which are to inform about SDR in a given item or scale only. These remarks regard also other methods (e.g. administration mode in the case of mixed-mode surveys, where certain items may be administered in different mode than others) showing that the person- *versus* item-level division is much less rigid than other categories in the SDR methods taxonomy. This division depends on measurement context and researchers' intentions towards particular method use.

Prevention		Remedy		
Intent	*Ability*	*External*	*Internal*	*Level*
instruction manipulation, e.g. warnings	time limits	survey paradata analysis, e.g. response time, mouse clicks	mixture models	*Person*
bogus pipeline (BPL) & similar methods		SDR scales & similar methods	person fit measures	
general measurement design, e.g. anonymity, stress and distraction minimisation		overclaiming technique (OCT)		
mode of administration		psychophysiological measurements, e.g. ERP, eye-tracker		
intrusiveness reduction methods, e.g. randomised response technique (RRT), unmatched count technique (UCT)		keying & weighting, e.g. Krueger's method	differential item functioning	*Mixed*
third party ratings	item formulation, e.g. subtle items, neutralised items	implicit measures, e.g. Bayesian truth serum (BTS), IAT	factor deletion/rotation methods	*Item/scale*
indirect questions, nominative technique	item response options, e.g. forced choice	content irrelevant bias indicators, e.g. bogus items, inconsistency scales, instructed response items	structural analyses, e.g. higher-order factors	

Table 1. Classification of SDR control methods. Based on Adair (2014), Dilchert & Ones (2012), Kuncel et al. (2012), Reeder & Ryan (2012) and author's ideas.

Literature review suggests that methods differ widely in their efficiency, flexibility and details of use. The subsequent paragraphs will briefly comment on their main characteristics, paying special attention towards their utility in accounting for non-deliberate SDR in low-stakes large-scale contexts. The review is not meant to exhaust the literature but rather to give food-for-thought by pointing to research lacunas and also to sketch the methodological context in which the overclaiming technique (OCT) was conceived and developed. It is also meant to show why many researchers perceived OCT as a good candidate for the golden standard method of SDR control (Paulhus et al., 2003; Randall & Fernandes, 1991).

4.5.3 Preventive methods: intent

Appeals/warnings are one of the methods counted as influencing participants' motivation to fake and most probably are the only preventive measure that is easy to apply in every research context and that is free from serious flaws.

Instruction manipulation

Manipulating task instructions is one of the proactive methods influencing participants' intent (motivation) to respond desirably (be it deliberate faking or non-deliberate self-enhancement) that seems effective and relatively easy to apply in practice. In this method special persuasive statements (oral or written) are embedded in task's instruction in order to blunt participants' motivation to respond desirably. Those statements can be positive, trying to persuade participants to respond honestly, by referring to moral principles (e.g. "results of scientific inquiry are important", "lying is not socially valued", etc.) or they can be negative, trying to influence participants through warnings (e.g. "fakers can be detected") often in the form of some consequences that may be held for those not giving "true" answers. Such consequences can entail loosing remuneration for research participation, exclusion from further studies or, in case of applied contexts, e.g. exclusion from further stages of job recruitment.

Pace and Borman (2006) proposed a taxonomy of manipulated instructions where negative warnings can inform about: a) detection methods ("fakers can be identified") or b) consequences ("fakers can be punished", "we will remove you from the remuneration list if you will provide dishonest answers"), on the other hand, positive warnings appeal to c) universal moral principles ("dishonesty is immoral"), participants' d) reason (e.g. "do not fake, as you will give wrong image of yourself, which may have adverse consequences, in example, you will obtain a job you are not really suited to") or they e) share information about the objectives of the study, trying to evoke participants' reciprocity in the form of candid responses. Additionally, this kind of instruction points to the value of scientific inquiry and hence often it is called "educational (appeal)". Another kind of positive appeal was introduced by Turcu (2011) who proposed appealing to the subjective norms of a given society (e.g. "students of our university do not lie"). Thus, there are two main kinds of negative instructions, warnings of detection and warnings of consequences, and few, relatively recently proposed and poorly researched, positive instructions often referred to as honesty instructions or appeals (Adair, 2014). Another method, engaging participants more but based on similar ideas as instruction manipulations, is the so-called solemn oath method. In this procedure, before doing anything else, participants are asked to sign an agreement that they will provide only true and honest answers. A vast majority of participants sign it without problems (Weaver & Prelec, 2013). As warranted by research evidence this method is moderately efficient but leads to more coherent answers, probably by reducing careless and inattentive responding (de-Magistris & Pascucci, 2014; Jacquement, Joule, Luchini & Shogren, 2013; Weaver & Prelec, 2013).

Literature review points out that relatively few research studies were devoted to the efficiency of warnings and appeals in response biases reduction. In a meta-analysis, Dwight and Donovan (2003) identified only 15 of such studies, whereas Adair (2014) found them only slightly above 20. Despite the low number of studies the results of the above cited meta-analyses are promising- Dwight and Donovan (2003) pointed out that manipulated instructions reduce faking by on average 0.23 of standard deviation. Even more promising results were presented by Adair (2014) and Zheng (2015), showing that this result may be underestimated and range even to 0.91 of standard deviation, depending on study design and trait measured. Nonetheless, the evidence on which type of instruction manipulation is the most effective is mixed. Some research suggests that negative instructions are the most effective, especially those accusing participants of dishonest responding or threatening them of consequences (Burns, Filipowski, Morris & Shoda, 2015), while other studies point out that appeals lead to more valid estimates of measured traits (Adair, 2014). Some results indicate that educational and subjective norm appeals may be the most efficient kinds among the positive instructions family (Adair, 2014; Zheng, 2015).

The above-cited results on instruction manipulation efficiency mostly come from personality assessment studies conducted in practical contexts of job-applicant selection or from studies simulating such contexts (Dwight & Donovan, 2003). Moreover, in most of those studies participants come from very specific subsamples, often highly-motivated to perform well in assessment (e.g. candidates for work promotion, military trainees, candidates for police officers; Dwight & Donovan, 2003; Eysenck, Eysenck & Shaw, 1974). These factors reduce external validity of manipulated instruction studies, as it is difficult to indicate how appeals/warnings would work in low-stakes contexts with participants not motivated to „fake good" but also not motivated to answer honestly, neither to pay specific attention towards instructions and questions as research situation would not bring any practical consequences to them. There is also a research gap that hinders to predict how this method of faking prevention would act if used not on personality scales, but on e.g. self-assessment of skills (e.g. educational skills, as it is done in ESS, PISA, PIAAC, etc.).

Hence, the applicability of these methods in low-stakes contexts is hard to evaluate due to lack of evidence. However, there are also other limitations to accepting instruction manipulations as a panacea to SDR in every research context. Warning participants of consequences is not possible in most low-stakes research situations as respondents in these contexts have literally nothing to lose and no practical repercussions can be held against dishonest participants (even if they can be identified somehow). This is especially true as warnings seem to bring some side effects (e.g. test anxiety, lowered motivation; Burns et al., 2015). In situations where participants are obliged or externally motivated to take part in a study (e.g. job selection, in-job assessment) it is probable that warnings' assets outweigh the drawbacks (lower motivation, test anxiety) but it is the other way around in most of the research situations where participants' motivation is of crucial importance to obtain high quality data (Meade & Craig, 2012). Thus, it seems that application of less-researched appeals is more promising than using warnings in low-stakes contexts. Apart from the limited utility in certain measurement situations another limitation of instruction manipulations is their restricted efficiency. First of all, participants may not pay attention to instructions and may even do not read them at all which is especially probable in self-administered modes. Available evidence points out that such omissions indeed happen, which cancels out all potential benefits of manipulated instructions[46] (Paulhus, Bruce & Trapnell, 1995). Thus, participants' non-compliance remains one of the most serious limitations of this method. Probably this drawback can be mitigated by method's developments such as solemn oath procedure. Additionally, this method may not be equally efficient towards various forms of positivity bias. It is evidenced that instruction manipulation is reasonably effective in high-stakes research occasions but will the method be comparably useful in other contexts? Nonetheless, there is evidence that appealing for honesty and diligence reduces incoherent survey behaviours thus enhancing data quality (e.g. Jacquement et al., 2013). It is not known, however, how instruction manipulations will affect non-deliberate positivity biases, e.g. self-enhancement. Nevertheless, relations between self-enhancement and inattentive responding (e.g. Paulhus et al., 1987) are all too obvious to ignore possibility to reduce them by instruction manipulations. Taking into account method's limitations it is worthy to remember that its advantages are also remarkable: it is an easy-to-use, cheap and flexible procedure that is already evidenced to be effective in response biases reduction.

Bogus pipeline

Method similar to appeals/warnings in that it also motivates participants to give true(er) answers is the bogus pipeline (BPL; Jones & Sigall, 1971). This method bases on convincing participants that their

[46] However, this research also bespeaks that ignoring or skipping instructions is not a common respondents' behaviour.

answers will be verified by an external device simulating a lie detector. In some cases the device is substituted by a sham biochemical analysis (e.g. collecting saliva samples) where participants are informed that it will be used to detect liars. Meta-analyses point to mixed results of BPL application (Krumpal, 2013; Tourangeau & Yan, 2007). The BPL has also some evident flaws such as serious ethical considerations (how far can we go in participants' deception?), large organisational effort (constructing a bogus lie detector), is impossible to apply in certain research contexts (telephone or web-based studies) and is very difficult to use in field studies. Due to organisational and ethical considerations this method is also completely unfeasible in large-scales studies and will not be commented further on. However, its logic, based on convincing participants that unwanted behaviours could be identified, is possible to use without constructing bogus lie detectors, e.g. through instruction manipulation, online warning application (cf. Burns et al., 2015) and even manipulating participants' honesty by priming (Rasinski, Visser, Zagatsky & Rickett, 2005)[47].

Measurement design

This category is in fact not a separate group of methods but rather a bunch of rules that should be observed in every measurement with human participants and that may have special importance in curbing SDR. Among such rules is avoiding or minimising of speeded responses, stress, tension, emotional arousal (if it is not part of experimental manipulation or a necessary element of measurement) and cognitive load- all these factors are known to increase SDR (Paulhus, 1991).

Another rule, especially important for positivity biases control, is assuring participants of their anonymity (Nederhof, 1985). In measurements where respondents are physically separated from each other, where they leave no identifying marks on their answer sheets and they do not pass over any identifying information about themselves (e.g. e-mail addresses), SDR is lower (Paulhus, 1991).

These methods alone do not guarantee sufficient SDR reduction but are necessary prerequisites in every measurement occasion if SDR control is to be possible.

Mode of administration

Mode of administration is a special part of measurement design often linked with substantial influence on research results and participants' behaviour (see e.g., Groves et al., 2009 for review). Early on it was noticed that participants are more prone to SDR when measurement was mediated through interviewer, be it in a classical face-to-face interview or in telephone interview, which both yielded more SDR than self-administered modes (Hochstim, 1967). This observation was later on confirmed by many studies, bringing firm conclusion that SDR tendencies are lowest in self-administered mode (Kreuter, Presser & Tourangeau, 2008; Krumpal, 2013; Nederhof, 1985; Tourangeau & Yan, 2007). However, the advent of computerised surveys brought new questions to the field: whether self-administered measurement on paper-and-pencil versions yield similar results as their computerised versions (including Web interviews)?

The evidence on this issue is mixed, as some studies point to higher SDR in computerised measurements than in paper-and-pencil ones (e.g. Lautenschlager & Flaherty, 1990; more examples in Paulhus, 1991), while others indicate no differences between the modes or even detect a reversed effect- lower SDR in computer-based measurements (review: Tourangeu & Yan, 2007). Dwight and Feigelson (2000) as well as Richman, Kiesler, Weisband and Drasgow (1999) noticed that differences between paper and computer versions diminished in later studies in comparison to earlier ones. This

[47] In this particular case respondents, prior to answering proper items, responded to a word-matching task that implicitly promoted being honest. Rasinski et al. (2005) reported that this method yielded less SDR in comparison to a control group.

result suggests that with the growing popularity and commonness of computers, as well as with improving computer abilities in subsequent generations (Burski et al., 2013), any differences between paper- and computer-based self-administered surveys disappear. This result is supported by the meta-analysis performed by Vispoel, Morris and Clough (2018) who voiced for full interchangeability between computerised and paper-based versions of the BIDR. Similar results, supporting no differences between modes in case of SDR scales and sensitive questions were also brought by Dodou and de Winter (2014) and Gnambs and Kaspar (2017). Interesting results were, however, presented by Koivula, Rasanen and Sarpila (2019) where differences between mail and web survey versions were found with mail version yielding more SDR. The authors indicate that such mode-differences are moderated by preferences towards certain mode by different participant groups, e.g. less educated and older respondents choosing mail over web and by ICT skills of respondents.

To sum up, it is evidenced that large-scale assessments relying on self-administered mode (e.g. PISA) are in less danger of SDR than these based on other modes (e.g. CAPI in PIAAC). Whenever conducting web-based surveys it is warranted to control participants ICT skills and Internet connection quality.

Intrusiveness reduction methods

Next group of methods are techniques that aim at increasing participants' comfort and convincing them that their responses are fully confidential. This goal is achieved through deploying special survey modifications in order to conceal responses from interviewers or to doubly assure participants that their responses cannot be tracked to them in self-administered modes. These methods are meant to reduce the subjective intrusiveness of measurement situation hence their other name of indirect methods. Most prominent among those techniques are randomised response technique (RRT; Warner, 1965) and item count technique (ICT) *alias* unmatched count (UCT; Raghavarao & Federer, 1979).

The former bases on giving participants two questions, one sensitive and one not, and a device generating random values (a die, a coin, etc.). Participants are instructed to use the device before answering every pair of such questions and informed that certain result (e.g. heads) obliges them to give a true answer on a sensitive question whereas any other result gives them freedom to answer truthfully the not sensitive question. As the interviewer do not know what is the result yielded by the random values device, the respondent can be assured that his/her privacy is not breached. There are many research results pointing that this method leads to more valid estimates (review: Lensvelt-Mulders, Hox, Heijden & van der Maas, 2005; cf. Hipsz, 2014), but there is also a large evidence that this method results in greatly overestimated values and hard to control measurement errors stemming from the method's complicacy (Holbrook & Krosnick, 2010a). Other studies also point that participants grow suspicious, often refuse to use the device and even purposefully do not obey the instructions in order to "hack" the whole system (review: Krumpal, 2013; see also Hipsz, 2014 for a detailed analysis of audio recordings of RRT measures). As evidenced by Coutts and Jann (2011) the method results in many "false no" results as respondents chose to use this safe option regardless to which item was pointed by the random device. Moreover, the method is difficult for many interviewers who struggle to perform it correctly and that respondents in self-administered modes also grapple with its correct use. Elevated measurement errors due to additional cognitive load in RRT on participants and pollsters was reported by Biemer, Jordan, Hubbard and Wright (2005). Furthermore, RRT is not successful in reducing respondents' suspicions regarding full confidentiality of their answers (Coutts & Jann, 2011) and may additionally increase participants level of test anxiety (Boeije & Lensvelt-Mulders, 2002). Further drawback of RRT is its very limited application in certain contexts, e.g. telephone or web-based studies (but see Coutts & Jann, 2011 and Höglinger, Jann & Diekmann, 2016 for recent developments in that matter).

The ICT/UCT technique consists of dividing participants into two groups. One of them answers to a set of questions, whereas the second one answers to the same set, but enlarged by one position containing an additional sensitive question. At the end participants from both groups do not yield their exact answers, but only a number of certain answers, e.g. how many times they answered "yes" to the questions on the list. By calculating mean differences between two groups a distribution of answers to the additional question can be estimated. Some research suggest that this method leads to more valid estimates than direct questions (e.g. Holbrook & Krosnick, 2010b). A big drawback of this method is a necessity to divide sample into two groups as it complicates research design and rises costs- to obtain same measurement precision as direct questioning a double number of participants is needed. Coutts and Jann (2011) compared ICT/UCT with RRT and direct questioning and obtained results pointing towards more valid estimates with the use of ICT/UCT.

A large problem of both these methods is that due to their specificity they do not provide information on individual respondent but only yield fraction of respondents that answered questions in a given way (e.g. indicated that they indeed cheated on a math test). RRT also has serious caveats regarding its efficiency and usability in most of the research contexts. A limitation typical to ICT/UCT is that it require larger samples to obtain similar measurement precision as direct questions (Cruyff, Boeckenholt & van der Heijden, 2016). Both methods were developed specifically to counter for deliberate forms of positivity bias, faking and blatant lying, and also answering intrusive, sensitive questions. It seems that their use is best restricted to these contexts and they cannot help much in countering non-deliberate biases. Nevertheless, ICT/UCT seems a viable option even for large-scales assessments where it can be used to measure sensitive topics, e.g. educational cheating, school aggression or bullying and inappropriate teacher behaviours[48]. RRT is not recommended to use due to large problems with procedure from both interviewers (if present) and respondents, as well as large fractions of false negative and false positive responses it generates (Höglinger & Diekmann, 2017).

Third-party ratings

This method, also called proxy-, informant-, other- or peer-reports, resides on collecting data about participant from other people, e.g. their partners, siblings or colleagues. Proxy reports have been mainly tested for personality reports and there is evidence for their good validity, less bias and measurement invariance with self-reports (Mottus, Allik & Realo, 2020; Oh, Wang & Mount, 2011). However, meta-analysis performed by Connolly, Kavanagh & Viswesvaran (2007) demonstrated that self- and peer-reports, though converging to a large extend (*r*s ~0.50-0.60), have predominant, substantive unique variance. These results indicate that this method can be treated only as an interesting research option and supplementary measurement[49]. Moreover, from the point of view of social psychology, the method entails important problems with self- *versus* other-perception, e.g. actor-observer differences[50] (Abele & Brack, 2013; Abele, Bruckmuller, Wojciszke, 2014; Abele & Wojciszke, 2007, 2014; Schwarz & Oyserman, 2001). These discrepancies can be additionally moderated by observer type (superior, peer, subordinate) and degree of acquaintanceship (family,

[48] Moreover, such use seems especially warranted as main drawback of ICT/UCT, using different item sets in different groups, does not constitute much of a problem in assessments using missing-by-design organization anyway (e.g. PISA). Further, the requirement of large sample sizes is also not a problem in ILSAs/NLSAs.

[49] McCrae (2018) showed that other-reports contained different content-irrelevant (method) variance than self-reports. It shows that informant reports contain both spurious and substantial variance that is unique in comparison to self-reports. Regarding the spurious variance it means that peer-reports are also biased but by different mechanisms than self-reports.

[50] For use of proxy reports in factual questions surveys see Dashen (2000), Schwarz & Wellens (1997) and Schwarz & Oyserman (2001). This research also confirmed that proxies use different response mechanisms and inferential strategies to answer survey items than would be used in self-reports.

relative, friend, etc.) (Connolly et al., 2007) and may introduce additional biases as demonstrated in a study by Cisłak (2013) whose title "All your boss can see is agency" tells the whole story about the nature of such biases. Moreover, other-reports are also susceptible to the regular biases present in self-reports (RS, C/IER, halo effects, etc.; Piedmont, McCrae, Riemann & Angleitner, 2000).

Apart from the above-mentioned, obvious caveats there are also two further issues with this method: a) so far it has been mainly tested on personality scales and its utility outside of these measures is largely unknown, b) if used to identify desirably responding participants in a large-scale assessment the method would yield enormous costs, at least doubling the project expenditure. The costs could be even higher if organisational effort of finding reporters matching in important characteristics, e.g. length and type of acquaintanceship, for each participant would be factored in. Hence, this method is not assumed to be a viable option for large-scale assessments due to its expensiveness and lack of verification in many research contexts apart from factual questions and personality reports.

Indirect questions

Indirect questions (IDQ)[51] are in a sense a mirror reflection of the third-party ratings- this time it is the respondent himself that is asked to report about other people. In this group of preventive methods participants are asked to report how "an average person" would answer to a given item (e.g. "what is the opinion of an average student in your school on cheating on exams") or what are the characteristics of this "average person" (e.g. "what is the level of mathematical abilities of an average student in your school") (Fisher, 1993). The logic of the method is based on reducing subjective intrusiveness of items if other people and not the respondent himself/herself are to being assessed. It is also assumed, basing on the false consensus effect (Ross, Greene & House, 1977), that respondent will project her own opinions and attitudes on others, hence yielding answers that would give an indirect window onto her own "answers".

Indirect questions underwent some verification that brought mixed results. Epley and Dunning (2000) demonstrated that reports on others were far less favourable than opinions on the self and that the self-reports were unrealistically favourable, whereas the indirect report were closer to reality (donations made in the donation paradigm: 2.44$ [self-report], 1.83$ [indirect question], 1.53$ actually donated). Similar conclusions were presented by Lusk and Norwood (2010) but Miller and Ratner (1998) provided evidence contradicting the results favouring indirect over direct questions (DQ). However, validity studies verifying IDQ over DQ are scarce. In one of the few examples Fisher and Tellis (1998) showed that IDQ did not correlate with the MCSDS, but DQ did, which the authors considered as a validity evidence in favour of IDQ. Notwithstanding, these results may be simply an effect of method variance as proposed by Jo (2000). The matter was thoroughly and innovatively settled by Jang (2017) who provided firm evidence that indirect questions do not lead to more valid estimates than direct questions when used as indicators of individual actual behaviour (in his study- charity donation). Moreover, he determined that the interaction between IDQ and DQ is important in predicting actual individual behaviour, bringing important evidence that IDQ and DQ may bring complement information. Jang (2017) also concluded that DQs always have higher predictive validity than IDQs and that DQs not always lead to enhanced self-reports. Furthermore, he suggested that projections theory does not seem to hold as respondents based their IDQs on other inferential processes. Interestingly, he demonstrated that using a "plea to exaggerate less" (type of honesty instruction manipulation) resulted in higher correlations between DQs and actual behaviour.

[51] Oftentimes RRT and ICT/UCT are referred to as "indirect methods". Hence, to differentiate between them and methods proposed by Fisher (1993) the Fisher's method is always referred to as "indirect questioning". Other term, though less used, is peer-prediction method (Miller, Resnick & Zeckhausen, 2005).

In another variation of IDQ respondents are instructed to report on a given person they know, which is known under the term "nominative technique" (NT), first proposed by Sirken in 1970s (e.g. 1970) and refined by Judith Droitcour (Miller) (1985). The method typically consists of two steps. In the first one respondents give number of their friends that did a certain thing, typically an illegal, threatening or unwanted behaviour (e.g. cheated on a math test). Second step takes place only if participant indicated at least one such person. In this stage respondents are asked to assess number of other friends that also know that the person named in the first step cheated on a math test. The second step is repeated for every person named in the first step. This procedure serves to correct for multiple reports of that one particular person (cheater) from many respondents (Krumpal, 2013). Typically, respondents are asked to answer also in-depth questions about one or every person nominated in the first step of NT. These in-depth questions may entail demographic inventory or items expanding knowledge about sensitive behaviour asked in step one (e.g. "when did he cheated", "how did he do it", etc.). NT is supposed to reduce the intrusiveness of questions by shifting the focus of interest from respondent to his/her friends or relatives. Moreover, it does not breach the anonymity of the nominated people as respondent is never asked to provide any identifying details (Lee, 1993).

However, NT is not free from serious methodological challenges. The method may seem simple and efficient, but in fact it may be cumbersome and frustrating for people that know numerous people that e.g. cheated on a test and are repeatedly asked about details regarding every one of this group (Droitcour, 1985). Participants' refusal to answer diligently in the second step, naming others that also "know", jeopardises the whole procedure and typically result in overestimates of cheaters due to use of inappropriate weights[52]. Obviously, participants mostly do not possess full knowledge about the others they describe, which is a common problem in proxy reports of which NT is not free either (Droitcour, 1985). Empirical verifications of NT are scarce, in one of the few studies, conducted by John, Edwards-Jones, Gibbons and Jones (2010), the method did not yield any results as participants simply admitted that they did not know about the inquired behaviours in anyone they knew.

Hence, scarce empirical validation of NT and many reservations towards its feasibility precludes from treating it as a viable option for an SDR control method in any larger assessment. On the other hand, IDQ is better tested but its validation results are not overly favourable. However, this method may be used, even in large-scale assessments, as a basis to some more sophisticated methods, namely Bayesian truth serum (BTS) commented on in one of the subsequent sections.

Summary of preventive methods (intend)

To sum up, the preventive methods group consists of a very diverse array of techniques, from simple and easy to implement (e.g. instruction manipulation) to awkward and cumbersome (e.g. BPL, RRT). Some of them may have their merits in research on sensitive matters but seem too complicated to use in general-purpose research like assessment of educational skills or personality (Krumpal, 2013). Regarding use in ILSAs/NLSAs it appears that instruction manipulation, ICT/UCT and, to some extent, IDQ are viable options for an SDR control method in large-scale assessments, with a caveat that utility of ICT/UCT may be limited only to sensitive questions. It is also noteworthy that instruction manipulations are most theoretically-grounded among the proactive methods.

[52] In order to estimate the number of cheaters in the population, the number of cheaters named in the first step is divided by 1 + the number of others that know. Satisficing in step two, a common problem of NT, results in overestimates of cheaters in the population due to too small number of "others that know" named in step two (Krumpal, 2013).

4.5.4 Preventive methods: ability

Other proactive methods influence not participants' intent to fake but their ability to do so. This is achieved mainly by manipulating items format (e.g. forced choice responses) or items transparency ("subtle" items, contextualised items, neutral items; Adair, 2014; Dilchert & Ones, 2012; Huber, 2017). Third possibility is to limit participants' time to answer the items, as it is theorised that faked responses take more time than honest ones (Dilchert & Ones, 2012).

Time limits

Imposing time limits in order to control for SDR is hardly researched at all, as the most recent meta-analysis identified only a handful of studies that used this technique (Adair, 2014). The results are mixed at best as most of the studies found no differences between measurements with time limit imposed (speeded condition) and without it (un-speeded condition) (Holden, Wood & Tomashewski, 2001; Robie, Komar & Brown, 2010; Robie, Taggar & Brown, 2009). Khorramdel and Kubinger (2006) showed that time limit had no main effect on self-report scores, but that it interacted with response format. However, the pattern of this interaction was different in different scales used and no clear patterns were identified. In a study conducted by Komar, Komar, Brown and Taggar (2010) time limit also did not significantly influence self-report scores but interacted with participants cognitive abilities, reducing SDR in participants with low cognitive abilities.

In this state-of-the-art the time limit method is not suitable for any practical application as it seems that further research in this path must be preceded by a careful theoretical explanation why time limits are to influence honest responding. Moreover, important practical issues are to be solved, e.g. what time limit serves best to unable SDR, how to incorporate inter-individual differences in reading speed and how to obtain baseline reaction times in order to set the most efficient time limit (Dilchert & Ones, 2012; Khorramdel & Kubinger, 2006). These steps are important as imposing time limits is perceived negatively by participants who reported higher tension and fatigue in speeded conditions (Roma et al., 2018; but see Komar et al., 2010 for neutral perception of time limits), which may induce satisficing and lower data quality in speeded conditions. Most importantly, however, the evidence from different paradigms indicates that, if anything, time limits seem to promote self-enhancement (e.g. Shalvi, Eldar & Bereby-Meyer, 2013; see also Paulhus, 1991), thus using this method to control for any type of SDR would be probably counterproductive.

Recently a new development of an old time limit idea was presented by Meade, Pappalardo, Braddy and Fleenor (2018) who presented a method called rapid response measurement (RRM). RRM resides on computer-based assessment where items are presented one at a time with short inter-item latencies. Respondents are asked to answer as quickly as possible. According to the authors the method retains psychometric properties in comparison to a traditionally-measured scale but participants were impossible to fake it when instructed to do so. RRM seems a moderately promising option for future research.

Item formulation manipulation

Another idea to undermine participants' ability to SDR is based on neutralising items, in order to strip them from any evaluative, desirable or threatening content. This is achieved by rewriting items to a new form, often called "subtle" or "neutral/neutralised", e.g. an item "Am the life of the party" was reformulated to "Prefer to be the central figure at a party". The new version is less evaluative and is predicted to yield lower scores and not induce SDR (Bäckström & Björklund, 2013; Bäckström, Björklund & Larsson, 2012). Another important aim of item reformulation is to make them less

transparent for participants in order to reduce possibility of inferring their purpose and tailoring reports to make a favourable impression (Adair, 2014).

Evidence on efficiency in SDR reduction in case of subtle items is favourable, as they lead to less favourably self-presentations simultaneously retaining good psychometric qualities of measurement scales (Adair, 2014), as evidenced by analysis of their effects on reliability, factorial structure, construct validity and criterion-related validity (Bäckström et al., 2009; 2012; 2014). Use of neutralised items even resulted in lower proportion of construct-irrelevant variance in self-report scales[53] (Bäckström et al., 2014).

Subtle items are hence hard to fake and evoke lower SDR tendencies but also convey certain drawbacks. First of all they need a considerable effort in piloting in order to develop items of truly neutralised content and retained psychometric properties. Moreover, sometimes subtle items present measurement problems, e.g. lower reliability, which means that a larger number of subtle items in comparison to transparent items needs to be used to obtain comparable measurement precision. Furthermore, neutralised items may be hard to use in cross-cultural research as they may pose exceptional difficulties in translation. The first drawback is not of a problem from the perspective of a large-scale assessment as they are normally supported by significant analytical background. The other two drawbacks present more serious obstacles as place is always short in packed LSAs' questionnaires and translational and cross-cultural issues are often in the very heart of organisational effort of every ILSA. Despite its advantages, usage of reformulated items may be also severely limited due to difficulties in employing them outside of personality assessment[54]. Of course this method cannot be completely efficient in certain research contexts, e.g. sensitive questions, factual items, etc. Hence, more research is needed on subtle items translatability and applicability outside of the personality measurement.

Item response format/categories manipulation

As the previous method entailed altering item content (wording) this method concentrates on the format in which response categories are presented. Most typically, self-report scales use Likert-type items in which respondent answers to every item separately. This kind of responding is easy to process for a respondent but also enables creating super-positive self-images by rejecting negative and embracing positive questions. One of the solutions invented to overcome this problem was changing response format into (multiple) forced choice (MFC). In this format respondent is to choose only one of the two items presented, e.g. "Chose adjective that describes you best: 1) practical, 2) imaginative" (Adair, 2014). The two (or more) options are to be equalised regarding their social desirability. In this way respondents are forced to choose among equally desirable (or undesirable) options which, in theory, results in yielding responses not distorted by SDR.

However, from the very beginning this method was burdened by measurement problems (Dilchert & Ones, 2012; Nederhof, 1985). The MFC design yields ipsative data that has peculiar and unwanted measurement problems in comparison to normative data: hard inter-individual comparisons (all respondents have similar scores), imposed score restriction (impossible to achieve an all high or an all low score) and inter-scales (inter-item) covariance structure, distorted criterion-related validity, inflated reliability coefficients and biased factorial structure (Brown & Maydeu-Olivares, 2013). These problems practically ruled out any serious application of these methods, however, the measurement

[53] Such variance is attributed to response biases: SDR or RS (Bäckström et al., 2014; Khorramdel & von Davier, 2014).

[54] Sometimes other techniques, e.g. presenting items in a randomised order, are described jointly with item neutralisation, however, they will not be commented here, as they fail to reduce SDR (Adair, 2014).

problems seem to be largely solved now with the use of IRT techniques to model MFC data (Brown & Maydeu-Olivares, 2011; 2013; Joubert et al., 2015).

Nevertheless, even with measurement problems solved, the MFC use to prevent SDR remains dubious. The evidence gathered to date suggests that MFC items are typically still biased by SDR in comparison to Likert-type scales and they are definitely fakable if respondents are told to do so (Adair, 2014; Dilchert & Ones, 2012; Pavlov, Maydeu-Olivares & Fairchild, 2019). Moreover, there is an indication than in case of the MFC data cognitive ability moderates data quality, with participants characterised by low cognitive abilities showing less SDR than individuals higher in cognitive functioning (Burns, Christiansen & Montgomery, 2005; Vasilopoulos et al., 2006). These differences are not normally found in rating scales.

Thus, it seems that forced choice responses utility is limited as they lead only to small or none SDR reduction in most of the contexts (Adair, 2014). There is also evidence that they may be useful in a strict contexts with very high stakes when deliberate faking threat is very real or when participants have especially strong SDR tendencies (Pavlov et al., 2019). It seems that such situation may occur in educational contexts where teachers are a group especially prone to SDR (Larson & Bradshaw, 2017; Wilhelm, Dewhurst-Savellis, & Parker, 2000; Żylicz & Malinowska, 2012). Other advantage of using MFC is that it can easily generate large numbers of unique items which would be especially welcome for large-scale assessments using computerised adaptive testing (CAT). However, MFC design is also notorious for being very time-consuming in piloting stages as every pair of items needs to be balanced on item desirability and other criteria (Dilchert & Ones, 2012). This property may be further discouraging from basing ILSAs on MFC measurement as the organisational effort in such programmes is usually exceptionally large.

Other manipulations of item response categories entail mainly changing number of rating categories in order to find option that works best in SDR reduction. Such comparisons are not often in the literature but some that were executed point to better properties of longer rating scales, e.g. 6-point, in comparison to dichotomous scales. Probably longer rating scales are less transparent and also enable to better differentiate between the hues of attitudes, opinions and traits. However, the difference is not large and is not present for very transparent scales (Khorramdel & Kubinger, 2006; Khorramdel, 2014). In general, continuously scored rating scales yield more valid results than dichotomously scored ones and SDR scales are no exception here (Cervellione,Lee & Bonanno, 2009; Kam, 2013; Stöber, Dette & Musch, 2002).

Summary of preventive methods (ability)

Preventive methods focusing on limiting participants' ability to fake yielded mixed evidence on their efficiency. The most promising option is taking item desirability under consideration when writing survey questions. MFC as SDR control method does not seem to rise to expectations but may be useful in cases when SDR tendencies (both deliberate and non-deliberate) are especially high.

4.5.5 Remedial methods: external

Survey paradata usage

Paradata is typically defined as administrative information on how survey was conducted, it can also contain information on response processes yielded by respondents (Couper, 2005). One of paradata examples that is researched as a potential SDR control method is response time (RT) analysis. In the above paragraph imposing time limits was commented on as not very promising method for SDR reduction (Adair, 2014). Here RT will be treated differently: as an additional source of information that

potentially enables SDR identification (see e.g. Kreuter, 2013 for more research on paradata and survey data quality).

First of all, there at least two competing conceptions of how SDR should be related with RT. The first of them, the Self-Schema model (Markus, 1977), proposes that processing information contingent with self-schema (a view of the self) should be faster than processing discordant information. Hence, according to this view SDR should yield longer RTs than "honest" responding as socially desirable image presented by SDR is not a true self-schema of respondents. The second theory, called Semantic Exercise model (Hsu, Santelli & Hsu, 1989), suggests exactly a reversed pattern, with SDR yielding shorter RTs than non-biased responding. This judgement is based on a believe that SDR is driven by cognitive processes that are much less complex than honest responding. This theory views SDR as a *de facto* satisficing[55], where rating decisions are taken basing on easy accessible and easy processable characteristics of items, e.g. their semantical features, whereas honest responding entails cognitively complex processes such as retrieving information from episodic memory. Another theory, proposed by Holden and colleagues (1992) and known as the Adopted Schema model, advocates that participants responding in a desirable way adopt an *ad hoc* constructed schema of an ideal self-image to which they compare their answers. As this self-image is a necessarily less complex one than real self-schema processing items in a desirable way is faster than "honest" responding. Another side theory proposes that deliberate SDR, e.g. faking, takes longer than honest responding as faking consumes additional time in response editing in which respondents tailor their answers in order to yield the wanted impression (Holtgraves, 2004).

Analysing the available evidence for RTs use as an SDR control method discloses that it is only in the beginning stages of organisation. Most importantly, there is still no consensus regarding which theory of chronometric properties of SDR should be accepted as a leading paradigm (Dilchert & Ones, 2012; Maricutoiu & Sarbescu, 2019; Xu et al., 2015). Moreover, neither of the RT-SDR theories is fully integrated with the available models of SDR/faking (e.g. McFarland & Ryan, 2000, 2006; Ziegler, 2011). Some convincing results in support of the Self-Schema model were presented by Akrami, Hedlund and Ekehammar (2007), however alternative explanations of their results exist and were not tested to date (e.g. Shalvi et al., 2013). In example, the pattern of results obtained by Akrami et al. (2007) resembles the relation between trait level and its desirability which is also an inverted-U shaped for most of the traits (Kuncel & Tellegen, 2009), hence their results could be driven not by self-schema comparisons mechanisms but by a simple heuristic of reacting to trait desirability. Moreover, there is quite a firm evidence that cognitive load elicits fast and socially desirable (e.g. self-enhancing) responses (e.g. Stodel, 2015; Xu et al., 2015), suggesting that the SDR-RTs relation may be much more complex and context-dependent than painted by Akrami et al. (2007) (see e.g. Maricutoiu & Sarbescu, 2019).

Regarding research practice there are still not enough data gathered to verify RTs utility as an SDR control method (Dilchert & Ones, 2012). Some evidence that RTs can be helpful in identifying participants responding in a desirable way was presented recently by Mazza et al. (2019). However, before RTs can be of any serious practical use a lot of methodological work is ahead of survey researchers as RTs are notorious for their measurement problems, e.g. low signal-to-noise ratio, high inter-individual variability, low test-retest properties, quantitative challenges due to their inherent characteristics (e.g. typical log-normal distributions), existence of specific effects (e.g. speed-accuracy trade-off) and moderating influences from many variables, including cognitive abilities, study design and item characteristics (Maricutoiu & Sarbescu, 2019; Paulhus & Holden, 2009; Wagenmakers, 2009; Wagenmakers & Brown, 2007). It is also worthy to note that efficient RTs use will require very precise

[55] Though they do not call it like that.

latency measurement as differences between cognitive mechanisms reside in milliseconds, hence this level of precision is needed also in survey research to make full use of RTs where latencies are now measured in seconds or even in minutes (Paulhus & Holden, 2009). Furthermore, any comprehensive RT analysis would require many item-level latencies in order to account for low methods' reliability (Paulhus & Holden, 2009).

Mouse-tracking analysis is another example of paradata use (e.g. Mazza et al. 2020), however, same reservations as in case of RT analysis apply to any kind of paradata, be it log-files, mouse-tracking and other forms of data. It seems that these methods are on a relatively initial phase of research and cannot be treated as sure methods of SDR control. However, as many large-scale assessments go online (e.g. PISA became fully computerised in 2015), paradata is going to be collected anyway. Thus, it is advisable to investigate best methods to use this potentially valuable source of information in response processes and response biases research. Inevitably, any advance in this field would also require theoretical integration of the now mutually incompatible models.

SDR scales

Next group of reactive methods to control for SDR variance is consisted of the so-called SDR scales or questionnaires. Other, often used names contain e.g. honesty, control, validity and lie scales, whereas unlikely virtues and infrequency scales are very similar concepts, which are treated jointly with the "classical" SDR scales here. The logic behind this approach is that individuals noting high scores on these scales will also yield untrue, distorted, socially desirable responses in self-report scales of substantial interest (e.g. math ability, math anxiety, etc.). The score of the scale can be used in quantitative data analysis, e.g. as a moderator, mediator or, most often, suppressor of a given empirical relation, e.g. relation between math ability as measured by self-report and an objective (cognitive) math test (Ganster et al., 1983; Piedmont et al., 2000). Such scales were also used in accounting for differences between interpersonal abilities measured by self-report and observers' ratings (Borkenau & Zaltkauskas, 2009) or intercorrelations of personality traits (spurious, inter-factorial variance; Holden, 2007).

The history of SDR scales dates back at least to 1930s. when Hartshorne and May (1930) initiated the long history of the „lie scales". Next major step was taken by Hathaway and McKinley (1951) who elaborated new lie scales with an aim to identify faking individuals in clinical research. Their works later on became known as the MMPI lie scales. These scales were used by Edwards (1957) who constructed the first widely recognised tool intended specifically to measure proneness to SDR. However, both the Edwards' scale and the MMPI lie scales were proven to have very limited scope of use and a narrow theoretical basis- both treated high SDR scores as a psychopathological symptom, moreover, many items had clear clinical connotations that lowered scale validity when used in non-clinical samples (Crowne & Marlowe, 1960). This situation was changed by new conceptions that appeared in 1960s. These new ideas called to cease treating SDR only as blatant lying or a psychopathological symptom. A prime example of these new approaches is the Marlowe-Crowne Social Desirability Scale (MCSDS) that became the main reference point for the next years of the SDR research. This scale treated SDR as a consequence of conformity and need of social approval, resulting in overly positive images yielded in self-reports, especially including denying common but undesirable behaviours, e.g. bad manners or undesirable thoughts (Crowne & Marlowe, 1960). The MCSDS won a huge popularity and was used worldwide, in many variants that included also adaptations to other languages (Cosentino & Solano,

2008; Holden & Passey, 2010; Sarbescu, Costea & Rusu, 2012; Siuta, 1989[56]; Verardi et al., 2010) and shortened versions (Clancy & Gove, 1974; Stocké, 2014; Strahan & Gerbasi, 1972).

Despite its popularity the MCSDS did not escape serious critique. The meta-analysis conducted by Beretvas, Meyer and Leite (2002) disclosed problems with reliability, especially in some age groups (cf. Neal & Carey, 2005). Other research questioned its adaptability to cross-cultural contexts (Verardi et al., 2010). The most serious charge was presented by Paulhus (1984) and Barger (2002) who pointed to the problem that the scale's factor structure did not correspond to its theoretical foundation. The MCSDS was designed as a one-dimensional inventory, whereas the empirical evidence pointed to a two-dimensional structure with correlated factors (Leite & Beretvas, 2005; Tao, Guoying & Brody, 2009). Apart from being in a serious contradiction to the Crowne and Marlowe's interpretation of need for social approval being the only construct underlying the MCSDS scores, this pattern of factor structure is not beneficial for any measurement tool as cross-loadings often impede clear theoretical interpretation of the results (Fabrigar, Wegener, MacCallum & Strahan, 1999).

Moreover, the MCSDS was soon considered to be based on an obsolete theory that only comprised a small fragment of SDR. The view of SDR changed to a multifaceted construct driven by several, often separate processes in which both individual's stable traits but also transient conditions of the measurement context play role (Hartshorne & May, 1930; Holden & Passey, 2010; Johnson & Van de Vijver, 2003; Kurzt, Tarquini & Iobst, 2008; Paulhus, 2002; Ziegler, 2015). The most commonly used contemporary SDR model postulates that the MCSDS measured not need for social approval, but it was a mixture of items measuring self-deceptive and impression management tendencies (Paulhus, 1984). Most of the MCSDS items are now classified as measuring agency management, type of the impression management (IM) construct (Paulhus, 2002).

The new two-dimensional model of SDR gained huge attention and empirical support, no wonder thus that also the new SDR scale constructed of the basis of it- the BIDR- was used widely (Paulhus, 1984;

[56] This is the most commonly known Polish translation of the MCSDS. However, other translations were also in use. To the best knowledge of the author of this dissertation no full adaptation of this scale to the Polish was ever done. Initial research on the scale adaptation was only presented by Siuta in 1989. In spite of that, the MCSDS is commonly used in research projects in Poland (e.g. Chachaj et al., 2006; Polczyk, 2005; Suszek, Fronczyk, Kopera & Maliszewski, 2018; Zinczuk-Zielazna & Słysz, 2018). Other translation was presented earlier than Siuta by Korzeniowski (1980), as stated by Izdebski (2007, 2013). Another translation was presented by Hipsz (2014). However, to my best knowledge, no psychometric characteristics of those two versions were ever presented. Another translation existed from early 1980s and was used by Zuber (1981) and by Wojciszke (1984). This time, some initial psychometric data was presented. A scale originally conceived in Poland but related closely to the MCSDS was prepared by Drwal and Wilczyńska (1980). This questionnaire, named Questionnaire of Social Approval (abbreviated KAS from the Polish name *Kwestionariusz Aprobaty Społecznej*), underwent full adaptation and a strict psychometric inquiry that confirmed its good psychometric qualities. The KAS correlates highly with the MCSDS (r~0.70; Drwal & Wilczyńska, 1980). Another SDR scale available for research on Polish samples is an adaptation of the SDS-17 scale (Stoeber, 2001) prepared by Fronczyk, Skrzyński & Cieciuch (2012). However, the report from this adaptation is not publicly available, which hinders its verification and further research on the Polish version of the SDS-17. Another scale, named Social Approval Scale (SAS, *Skala Aprobaty* Społecznej) was constructed by Żylicz and Malinowska (2012). This scale is based on items driven from the KAS, the MCSDS, the BIDR and a scale on morality and social values conceived by Żylicz. The SAS was used in one large-scale research project on a representative sample of the Polish teachers, but the scale results and items were not made available for the public. This makes the KAS the only Polish-language SDR scale that is available in the public domain and has proven psychometric and substantive properties. Paradoxically, despite at least four commonly known BIDR translations into Polish, one of them presented by Hipsz (2014), the other used by Zinczuk and Draheim (2009), this scale did not enjoy great research interest in Poland (see Izdebski et al., 2013 for an isolated example of its practical use). The first formal adaptation of both the BIDR and the MCSDS based on new, revised translations is still underway in the research project led by the author of this dissertation.

Paulhus & Notareschi, 1993). Out of the two SDR dimensions it was the IM, purposeful and conscious distortion of self-reports, that was perceived as a bigger threat to the validity of research results. The other bias, self-deception, was assumed to be a minor threat (Paulhus & Vazire, 2007). However, further research brought weighty allegations against the accepted classification and role of SDR scales in self-report research. Uziel (2010) showed that IM, as measured by the BIDR, did not correlate with systematic cheating and being insincere (cf. Li & Bagger, 2006; Ones & Viswesvaran, 1998). Uziel's research points that high IM scores (either from the MCSDS or from the IM subscale in the BIDR) do not predict distortion of self-reports. Instead, such high scores can simply indicate a well-adjusted individual, agreeable, non-impulsive and with good self-control. Thus, Uziel challenged the validity of SDR scales (the IM-BIDR and the MCSDS) to measure distortive tendencies in self-reports, moreover, he also presented evidence that individuals with high scores on the IM scales may in fact yield more valid, corresponding to reality self-reports as confirmed by higher correlations with peer-reports in case of high IM-scoring group in comparison to low IM-scoring group (Uziel, 2014). Uziel's research not only questions the validity of SDR scales as such but also shifts the attention from IM to self-enhancement as a main SDR-related threat to the validity of self-reports (at least in low-stakes contexts).

These results point to the crucial objection against using SDR scales to measure distortive tendencies in self-report at all[57]. Many studies showing that SDR scales scores failed to act as a moderator, mediator or suppressor of response biases of many kinds, e.g. empirical relations between self-report and verifying criteria, e.g. IQ test or proxy-report (Franzen & Mader, 2019; Leising, Scherbaum, Locke & Zimmermann, 2015; Piedmont et al., 2000; Smith, 1997; Uziel, 2010; meta-analysis: Ones et al., 1996). Li and Bagger (2006) determined inability of validity scales to detect spurious variance, while Huang (2013) showed that SDR scales explained only minor spurious effects on substantive relations between self-esteem and academic performance. Similarly, Lönnqvist, Verkasalo and Bezmenova (2007) claimed that SDR scales offered no help in identifying fakers under neutral instructions. They also failed to find any moderating effects of SDR scales. Moreover, in some cases employing scores adjusted on SDR scales resulted in reduced validity in personality scales (Piedmont et al., 2000). It is also worthy to compare the results by Zettler, Hilbig, Moshagen and de Vries (2015) where low, instead of high, scores from the IM of the BIDR were related to cheating in an online game. These results point not only to zero, but also adverse, counterproductive results of using SDR scales.

Two possible explanations of this failure were proposed: first, that SDR scales do not measure tendencies to present oneself in socially desirable, overly positive light and second, that the scores of the scales are distorted itself, e.g. by other response biases, e.g. response styles or careless responding. The former probability is known under the substance *versus* bias debate (McCrae & Costa, 1983). In this line of thinking SDR scales are believed to measure substantial, trait-like variance, "overlooked aspects of content variable" , as put by Sackeim and Gur (1978). Many interpretations of this overlooked aspects were proposed. In example, Uziel (2010, 2014) stated that SDR scales measure motivation to include socially desirable norms in one's life and behaviour which he called interpersonally oriented self-control. De Vries, Zettler and Hilbig (2014) found correlations between the HEXACO personality inventory and the BIDR: the SDE subscale correlated positively with Extraversion and negatively with Emotionality and Conscientiousness, whereas the IM was related to Conscientiousness, Agreeableness and Honesty-Humility traits. Moreover, Guo, Liu, Wang, Li and Gao (2019) determined that high scores on SDR scales were predictive for good job-related performance,

[57] In example: "In fact, these problems are important enough to indicate that SD scales are inappropriate both for applied purposes and for the purposes of researching the mechanisms and processes underlying faking behavior." (MacCann, Ziegler & Roberts, 2012)

good mental health, including boredom enduring, adjustment after graduation, buffered social stress and increased stress-coping resources. Van der Linden, Dunkel and Petrides (2016) also related the IM scores to job-related performance, attainment of social goals and social effectiveness. It seems that high SDR scales, especially obtained from the IM subscale of the BIDR, are related to honest behaviours and good social adjustment. It seems that such scores are indicative of a good, desirable personality, not of response bias (see also Ones et al., 1996; Piedmont et al., 2000).

The latter explanation, SDR scales being distorted by other response biases themselves, was voiced e.g. by Becker (1976). Many studies pointed out that the SDR scales can, and often are, faked themselves (Furnham & Henderson, 1982; Marselle, 2014; Moorman & Podsakoff, 1992; Viswesvaran & Ones, 1999). Such scales may be specifically susceptible to being faked (or distorted by other response biases) due to their high transparency which enhances participants abilities to distort self-report results (Dilchert & Ones, 2012). An interesting example of such problems was provided by Paulhus (2002) who admitted that participants often have problems with the self-deception denial (SDD) subscale of the BIDR as they consider its items as intrusive or even offensive. Moreover, it is possible, due to the specific construction of the SDR scales, that their items have higher desirability that any other items used in the studies, which would mean that the SDR scales are more prone to distortion than any other research instruments in the surveys. What is more, the SDR scales are deemed easily coached, meaning that participants can learn how to cheat them in order to obtain desirable scores, which greatly limits use of this method in practical settings (Alliger, Liliendfeld & Mitchell, 1996; Hurtz & Alliger, 2002; Miller & Barrett, 2001; Robie, Cartina et al., 2000). Other problems related to SDR scales were pointed by DeMaio (1984) who noticed that the process in which many of the scales were build is questionable as respondents had serious problems in assessing the desirability of items on the initial stages of building SDR scales. Furthermore, SDR scales often contain culturally-specific desirable behaviours and are thus difficult for cross-cultural research which additionally hinders their theoretical use in large-scale assessments (Lönnqvist et al., 2007).

However, despite the amassing evidence of problematic SDR scales use, they still remain the easiest and most convenient way of measuring (alleged) SDR tendencies. Oftentimes, due to ethical, organisational or methodological issues other methods are not possible to use, thus leaving the validity scales as the sole remedial method of controlling SDR in many research endeavours (Krumpal, 2013; Nederhof, 1985). Moreover, the substance *versus* bias controversy continues, as some studies point to successful implementation of validity scales for SDR control (e.g. Berry, Page & Sackett, 2007; see also Steger, Schroeders & Wilhelm (2020) for more examples). Hence, the SDR scales utility calls for a thorough, meta-analytic study which would comprehensively describe lie scales properties as an SDR control method depending on sample, design, criteria and scales used. Insufficiency of the presently used scales does not preclude efficient use of other scales, e.g. dark personality, self-orientation or competitive worldviews inventories (Roulin & Krings, 2016; Van der Linden et al., 2016).

Overclaiming technique (OCT)

The overclaiming technique was linked with very high hopes of being an efficient and easy-to-use method of both deliberate and non-deliberate SDR control (Paulhus et al., 2003; Philips & Clancy, 1972; Randall & Fernandes, 1991). It is comprehensively reviewed in the subsequent chapter.

Psychophysiological measurements

These techniques join information from human's physiology and cognition by measuring biological substrates of psychological (socio-emotional and cognitive) processes. Among prominent methods used in this field are: a) brain activity measures- imaging, e.g. functional magnetic resonance imaging

(fMRI) and electroencefalography (EEG), b) eye activity measures, e.g. pupillometry and eye-tracking, c) cardiovascular activity, d) muscle activity (myography) or e) electrodermal activity.

So far these techniques found significant use in the SDR research despite their large cost and organizational requirements (see subchapter 3.4.3 of this work). In this subchapter eye-tracking studies will be further commented on due to their notable role in accumulating evidence on survey responding in the recent years (e.g. Galesic et al., 2008), and even establishing the method as one of the essential pretesting tools (Lenzner, Kaczmirek & Galesic, 2011). Nevertheless, eye-tracking research concentrated specifically on SDR is still rare. Three examples come from faking studies in which eye-tracking helped to identify strategies used by participants in fake good conditions. It was evidenced that faking respondents concentrate their gaze mainly on extreme response options, but also read items in a less organized way with lower number of fixations (Muller, Schiepe-Tiska & Strohmaier, 2017; van Hooft & Born, 2012; Xu et al., 2015). In a preliminary study also using an eye-tracker Kaminska and Foulsham (2013) suggest that SDR may be a by-product of satisficing in which respondents read questions quickly and concentrate on extreme options directly after reading them. Eye-tracking seems a promising option for future studies on response biases, especially that it is readily combinable with paradata measurement in computer-based studies.

Another evidence specifically concentrating on SDR was brought by event-related potential (ERP) studies where the so-called P300 component (P3b variant) was found to have lower amplitude in participants scoring high on SDR scales (De Pascalis, 1993; Robinson, 2001). This pattern of results could suggest that high scores on such scales are driven by inattentive responses or responses driven by high cognitive load (P300 amplitude is lower in high cognitive load conditions, large distraction or force multitasking pressure; see Kok, 2001 and Linden, 2005 for reviews). Such pattern of results should be confirmed by further studies.

Psychophysiological methods will probably never be a core SDR control method, at least not for ILSAs, mainly due to their costs and organizational requirements (carefully prepared laboratory settings). However, these methods already proved very useful in explaining response biases mechanisms and their further use in this area along with an integration with other methods present in the field (e.g. paradata use) are much awaited and even necessary steps.

Keying and weighting techniques

This group of methods comprises of two related techniques. The first one is keying, which is an empirical scoring of items regarding their properties attained in measurement, e.g. predictive power or criterion-related validity (Mitchell & Klimoski, 1982). The second one is weighting, which is based on ascribing additional values (weights) to items in order to create additional indicators of response bias. Normally these weights are, like keys, also empirically-driven. One of few accessible in the literature weighting examples was presented by Krueger (1998) who used participants' item desirability ratings as an indicator of their self-enhancement tendencies. In this method participants completed a type of personality questionnaire in the form of list of adjectives ("Does the adjective describe you?") and then rated desirability of the items used in the measurement (e.g. easy-going). The obtained measure is then correlated with participants ratings on personality questionnaire. In the final step this correlation between personality and item personal desirability rating is adjusted on item social desirability ratings, based on averaged ratings of descriptiveness and desirability of the whole group. The resulting partial correlation is treated as an SDR index. The method possesses certain appeal (face validity?), however remains largely untested, similarly as other keying and weighting methods (Kuncel et al., 2012). Paulhus and Holden (2009) pointed to two drawbacks of the method: a) it takes precious time and space in questionnaire as respondents have to rate each item twice, b) the

order of ratings seem to influence the results. Problems related to assessing items' desirability reported by DeMaio (1984) rise additional doubts whether this method truly is capable of bringing any useful information. Before these methods would be used in large-scale assessments they need to undergo a thorough investigation of their true characteristics. However, it seems unlikely that any method based on doubling the number of items would become popular in the era of short scales domination.

Implicit methods

Bayesian truth serum (BTS) introduced by Prelec (2004) is one of the examples of implicit or indirect measures of SDR. In this task participants are asked to do two things: a) respond to a question and b) assess how other people will answer it (e.g. what fraction of the whole population will chose this option). The method resides on the false consensus effect and is based on using both answers in calculating individual scores where respondents are rewarded for correct predictions on how others will respond and for providing answers as close as possible to the answers predicted by the whole group. The method received some empirical testing and provided promising results as groups warned of the BTS calculation yielded less desirable and more honest answers (Frank et al., 2017; Weaver & Prelec, 2013). However, there are serious caveats regarding the methods' utility as just the same results could have been achieved if respondents simply reacted to warning of a "lie detector", similarly as if it was a BPL or instruction manipulation (Kuncel et al., 2012). The method needs further testing before anything decisive about its efficiency could be said. It is worthy to note, that this method is based on similar predictions as IDQ commented above and similar caveats towards its validity apply (Jang, 2017).

Gregg (2007) proposed using Implicit Association Test (IAT), a cognitive task widely used in experimental social sciences (Greenwald, McGhee & Schwartz, 1998) that resides on measuring implicit preferences towards certain objects. Gregg and Klymowsky (2013) noted method's pluses and minuses, however its utility for SDR research is still not known. Although the method is often used to measure racial prejudices and other undesirable attitudes and opinions it was never, to the best knowledge of the author, compared with measures and tools used in survey research. However, IAT seems a promising option for research investigations on SDR. Nevertheless, its use in large-scale assessments would be probably very problematic due to large time requirements of the method.

Construct-irrelevant bias indicators

One of the main disadvantages of SDR scales is that they capture trait variance, apart from bias variance, hence the idea to construct similar scales but this time measuring only response bias. The scales were named trait-free indicators, representative indicators of response styles (RIRS; Weijters, Schillewaert & Geuens, 2008) or response inconsistency scales (Piedmont et al., 2000). The author of this thesis named them "construct-irrelevant bias indicators" (CIBI) in order to underlie that such indicators need to be unrelated to the measured construct to be effective (cf. Ferrando, 2005). This name also alludes to the construct-irrelevant variance (CIV) term which the CIBIs are precisely meant to measure.

This category entails a large and internally diversified group of techniques. Instructed response items (IRIs; e.g. "If you read this mark "9" in this item"), bogus items (e.g. "I was born on Mars") and instructed manipulation checks (these items typically consist of a short instruction and an activity respondent have to do, e.g. provide an open ended response) (Curran, 2016). These items serve for C/IER identification and do not offer much help as SDR indicators.

However, two other forms of such items could be helpful. The first of the two is called "response inconsistency scales" and is formed of pairs of related items, e.g. "Normally I seek quiet places" and "I like being part of a large crowd". Paired items are related to each other in order to enable measuring inconsistent responses (Dilchert & Ones, 2012). In the example above agreeing to both items would be a potential signal of inconsistent responding. This method is normally used to screen for C/IER and RSs but theoretically may be set up to indicate SDR (see Piedmont et al., 2000). In order to achieve that item pairs would have to be equalised on desirability and paired in such a way that desirable responses to both of them would be logically contradictory. Such a scale would not necessarily be trait-free (Weijters et al., 2008) but would have to be of a significant length in order to guarantee needed precision of measurement (Dilchert & Ones, 2012). Existing examples of similar scales consist of more than 40 items, hence their utility is constricted mainly to high-stakes settings.

The second CIBI method of some potential to measure SDR is called self-reported honesty (Meade & Craig, 2012), diligence scale (Curran, 2016) or seriousness check (Aust, Diedenhofen, Ullrich & Musch, 2013). This is a very simple technique based on asking respondents direct questions regarding their honesty and attentiveness during survey. Such questions can be e.g. "Should your data be used in analysis? Yes/No" or "Did you answer honestly?". Despite its simplicity the method seems to work reasonably well in identifying low quality data (Curran, 2016; Meade & Craig, 2012). However, two caveats apply should this method be of any value to indicate SDR: a) check should entail more than just one item to achieve acceptable reliability and avoid false positives, b) its content should be more directed to honesty and desirability and less to motivation or diligence.

Both response inconsistency scales and diligence scales need preparatory work before they could be attested as SDR indicators, however these methods offer some possibility to become useful if only their length could be held at moderate item numbers.

Summary of external remedy methods

Remedial methods in this category are based on collecting additional measures, that are not related to the construct measured and their sole aim is to indicate SDR. These additional measures differ in their efficiency, ease of use and flexibility to be applied in different modes and contexts. SDR scales was a perfect candidate for gold-standard SDR control method, however, recent evidence amounted to a firm and consistent knowledge that these measures should not be used in this role. Hence, OCT is a second best hope to be an easy-to-apply but valid indicator of SDR.

Moreover, most of these methods is only in the initial phase of testing and all of them seem to lack verified theoretical underpinnings. Such a link between theory and practice is much needed in this context, in order to avoid misinterpretation of the collected indicators.

4.5.6 Remedial methods: internal

Mixture models

This method resides on identifying latent subgroups among the participants that took part in the measurement. The technique accounts for shared patterns in the data and can discern between groups of participants of different properties even if no grouping variables were explicitly stated which seems to be a blessing for SDR research where generally it is not known who yielded biased responses and to what extent.

Mixture models are widely used in social and life sciences but are only in initial phases of testing in response biases identification (Zickar & Sliter, 2012). However, some very promising evidence exists as Leite and Cooper (2010) have successfully used this method to differentiate between participants

responding desirably and honest responders. Other applications involve study of Meade and Craig (2012) who used mixture models to identify C/IER and analyses conducted by Khorramdel, von Davier and Pokropek (2019) who used this technique to discern different RS groups.

It seems that method offers a promising avenue to identify subset of respondents responding desirably, moreover, it allows to differentiate between different forms and patterns of SDR if such can be found. Mixture models is an option especially suited for large-scale assessments as they do not need any additional indicators and can offer an interesting insight into SDR on their own. What is more, the abundance of data normally collected in LSAs can be readily used to interpret groups emerging from mixture models. Method's large requirements regarding sample size are easily met in case of large-scale assessments.

Person fit measures

While mixture models are meant to identify latent groups in the data, person fit measures (perfits; PFMs) are devised to select outlying individuals. Normally, perfits are used to find aberrant patterns in the data, e.g. respondents who do not follow the instructions, and have been applied with promising results in detecting C/IER (Conijn, Emons & Sijtsma, 2014; Curran, 2016; Meade & Craig, 2012). The logic of analysis is similar to conducting regression diagnostics and detecting outliers. However, in this case person fit measures indicate discrepancy from an underlying IRT model for a given dataset, e.g. self-report scale.

Despite some successes in the C/IER research perfits were so far less successful in the SDR field (Zickar & Sliter, 2012). A set of simulational and empirical studies conducted by Conijn and colleagues showed that perfit measures performed best for random responding, but had less success in accounting for ERS or SDR in the data. Moreover, eliminating participants flagged as outliers brought only little gain in model fit and criterion-related validity. The method seems to be efficient in eliminating only certain types of outliers but does not help to identify truly influential cases, elimination of which could give substantial improvements in data quality. Nevertheless, the PFMs were able to outperform other SDR control methods, e.g. inconsistency scales, in a preliminary research (Conijn, Emons, De Jong & Sijtsma, 2015).

PFMs also have certain drawbacks. First of all, they are based on an underlying IRT model, most often graded response model (GRM), so any departures from models requirements (e.g. local dependency of items, multidimensionality) may result in deteriorated perfits efficiency. Although this is theoretically a crucial concern, empirical data seem to suggest that PFMs are quite robust to model violations (Conijn et al., 2014). Secondly, the method needs quite lengthy scales, of 20 items or more, to work properly. Moreover, items should vary in their fakability and desirability as method seem to be less efficient in identifying stylistic responses. Thirdly and finally, there is a large need for a theory that would enhance making sense of the PFMs results, as for now many aberrant patterns are hard to interpret.

Nevertheless, the method is an easy to use and cost-efficient technique that can be merged with other remedies, e.g. mixture models and DIF in order to account for various types of response biases present in the dataset (Conijn, Sijtsma & Emons, 2016).

Differential item functioning (DIF)

This technique allows to compare item properties between two groups of participants that differ only in one trait, but stand the same on underlying latent ability. In theory it is an ideal method for SDR research as it should enable detailed comparisons between honest and responding desirably groups.

However, the criterion that is used to form two comparison groups has to be known prior to analysis and collected during measurement. Nevertheless, DIF seems as a good method to expand knowledge on SDR by providing detailed item-level data. Its main drawback, necessity of a valid dichotomous criterion, is mitigated in LSAs as in these assessments numerous measures are collected anyway, so there is plenty of candidates for a criterion and performing DIF analysis may seem a viable option. It is worthy to note, that more research is needed to determine what measures possible to collect in LSAs could be used as a valid DIF criterion for an SDR identification (Zickar & Sliter, 2012). The DIF's equivalent for a scale-level analysis is differential test functioning (DTF). Measurement invariance analyses can also be used to determine similarity in model estimates between the pre-defined groupings.

Factor deletion/rotation

Paulhus (1981), basing on an observation of Messick and Jackson (1972) that the largest (principal) factor emerging from factor-analysis of the MMPI data was very highly correlated with item desirability ratings, conceived that deleting this factor from further analyses would eliminate SDR variance from the data matrix.

The method proposed by Paulhus (1981) was heavily critiqued by Borkenau and Amelang (1985) who provided empirical evidence that deleting such factor diminishes content saturation in resulting factor scores. Moreover, the adjusted scale scores were characterised by lower construct and structural validity in comparison to the unadjusted scores. The researchers firmly discouraged from using factor deletion technique. In fact, this method was quickly abandoned.

Apart from lack of efficiency the method also had one inherent drawback- it needed a valid SDR criterion to be collected with the target scale in order to serve as marker variable. Otherwise, there was no possibility to confirm that the emerging principal factor was indeed representative for SDR variance. As suggested by Borkenau and Amelang (1985), the SDR factor was not always the largest factor and even was not always present in the data. Moreover, its character seemed to depend from the scales' content, which limited method's use due to low inter-sample generalisability (see Paulhus, 1984). It appeared that this method confused item desirability (item characteristic) with SDR (actual survey behaviour).

Higher order factors

Higher-order factor method entails modelling variance common to subscales or even separate scales used in one measurement occasion. Oftentimes such common variance emerges as a so-called method variance (McCrae, 2018; Podsakoff et al., 2003) but can also represent response biases, e.g. RS (Khorramdel & von DAvier, 2014) or SDR (Podsakoff et al., 2003). Such common variance may provoke spurious correlations between theoretically unrelated (sub)scales or suppress true relations, thus lowering scales' construct validity. Another effect of common variance presence is spurious scale multidimensionality which causes both psychometric and substantial problems (Rauch, Schweizer & Moosbrugger, 2007).

Schmit and Ryan (1993) presented a study where typical factor structure of a Big5 questionnaire did not emerge in a sample motivated to yield favourable impressions (applicants for a job post). In the applicant subsample, in comparison to the non-applicant one, inter-factor correlations were observed, that were accounted for by a higher-order factor (HOF, also general factor, GF) in order to establish an adequate model fit. The authors interpreted these unexpected cross-factor correlations to SDR.

Many subsequent studies also found such a structure, but the interpretation of the general factor emerging is mixed (Pelt, Van der Linden, Dunkel & Born, 2019), as some researchers attribute it to response biases (method variance), whereas some perceive it as an emanation of substantial variance.

Dunkel, Van der Linden, Brown and Mathes (2016) contend that the emerging higher-order factor can be attributed to three separate sources of variance: social effectiveness, positive self-evaluation and SDR, the first two being sources of valid trait variance. On the other hand, Anglim, Morse, de Vries, MaCCann and Marty (2017) found that higher-order factor explained two-thirds of the difference between applicants and non-applicants, thus pointing to the bias (SDR, faking) interpretation of the higher-order factor. However, Pelt et al. (2019) and Dunkel et al. (2016) are more inclined to substantial *versus* bias explanation of HOF. Overall, the nature of such patterns is still elusive and validity studies are much needed, especially as some previous studies lacked ecological validity by concentrating on instructed faking studies where the role of fakers was played by college students (Pelt et al., 2019).

Bäckström (2007) provided some validation on this topic by obtaining correlations between HOF and the IPIP equivalent of the BIDR, which would suggest a bias nature of the higher-order variance. However, this validation still does not respond the substance *versus* bias question as the SDR scales are known to contain large substance variance themselves.

Reise, Kim, Mansolf and Widaman (2016) provided a detailed analysis on why higher-order models seem to offer better fit than unidimensional models in the presence of biased responses. The authors used a newly-introduced method- iterated reweighted least squares (IRLS) that allows to combine individual fit measures with many structural models, including various higher-order factor models (e.g. bifactor model). The method was used on a large database and showed that only a handful of participants (3% of the sample) was modelled better with a HOF model than a unidimensional (more parsimonious) model. Interestingly, a substantial part of the sample (11%) yielded such inconsistent responses that they did not fit to any model. As evidenced by a detailed data exploration the bifactor model helps to model stylistic responses, e.g. acquiescence (ARS). However, neither bifactor nor unidimensional model helped to model blatantly aberrant response patterns, e.g. straightlining or very low response variability (using the same two or three response categories throughout the whole scale). Moreover, there were still many aberrant patterns in the data for which an explanation was difficult to find.

Nevertheless, modelling higher-order variance and subsequent partitioning it to specific variance sources is a promising, yet under-researched SDR control method (McCrae, 2018; Pelt et al., 2019). A serious limitation of the research to date was that almost all of the results were drawn from personality or self-esteem scales, thus narrowing the research contexts in which this method was used.

Summary of internal remedy methods

Internal remedy methods are on the first glance very attractive as they can be used after the data was collected (e.g. on a secondary dataset) and typically do not require preparatory efforts. Moreover, they can be used as both confirmatory and exploratory techniques which gives a valuable flexibility (Conijn et al., 2016). What is more they are time- and cost-efficient as they necessitate only additional analytical exertion which is easier to provide than additional respondents' effort. Another advantage of these methods is that they easily combine with themselves (e.g. Conijn et al., 2015, 2016) mutually aiding interpretation as they can serve as criteria for each other. In example, perfit measures can be combined with mixture models and either perfit values or latent class membership could be used as a criterion to establish groups for a DIF comparison. This possibility was exploited by Reise and colleagues (2016) who combined logics of higher-order, person fit and mixture models in a newly

developed method of analysis. Certainly, more of such developments proposing innovative methods that allow to combine strengths of the existing models are most welcome.

Nonetheless, these methods are still under-researched which means that they still may need collection of additional measures that would serve as criterion in order to aid their interpretation. Moreover, theoretical advancements would greatly ease usage of these methods, as so far possibilities to generate models and measures seem infinite but incisive interpretations are in short supply.

4.5.7 SDR control methods: summary

The above review was meant only to reframe rather than exhaust the literature on response biases control methods so inevitably it is very selective. The main point of the review is that none of the methods used so far achieved the "golden standard" status and still there are no easy answers for difficult questions on how to account for the spurious measure caused by response biases, including the overly positive bias of self-perception. It is warranted to conclude that the research on methods was engulfed by pragmatic aspects and, consequently, achieved few theoretical advancements on the nature of the biases. Moreover, the research to date often entangled SDR, response styles, careless responding and other biases; the examples of studies that aimed to investigate relations between them are scarce (e.g. Grau et al., 2019; Ludeke & Makransky, 2016; Pokropek, 2014).

Most importantly, however, the research to date was very focused on the response biases in the "old", "classic" sense: as effects of social desirability or conscious, deliberate manipulation of self-reports, e.g. to achieve some gains (faking research) or to evade answering to intrusive questions. This research corresponds to, e.g. Goffman's or Blau's theories on impression management but it does not offer much in case of the new framework of positivity bias, both in case of theory as in case of methodology. There is a great need to develop new methods that would account for response biases also in other measurement contexts, namely in low-stakes, non-intrusive, non-threatening measurement occasions, which are common situations where response biases in the "new" sense emerge: results of unconscious, non-deliberate, motivated processes.

It is also advisable to think about the requirements that efficient SDR methods should fulfil. Kuncel and colleagues (2012) named a few of such specifications: high sensitivity, low rate of false positives, capturing only trait-irrelevant variance, resistant to coaching and instruction. This list is slightly tilted towards the needs of applied research (e.g. job selection) and can be supplemented by stating the most fundamental prerequisites of SDR control methods for any kind of research with a special focus on large-scale assessments. Such techniques should be time-efficient, especially on the side of the respondent (client). This is especially important in large-scale assessments as their questionnaires are normally packed with scales and there is small possibility of adding lengthy, time-consuming tools. Moreover, such method should be of course efficient, namely it should be evidenced to enhance self-report validity by accounting for construct-irrelevant variance. Furthermore, such method should be also easy to use, cost-effective and flexible, ready to use in various research contexts, designs and modes. It is also important to remember that the "golden-standard quest" is now more of a rhetorical figure than a true search for the best method. The evidence amassed so far points that SDR is not a zero-one phenomenon and it seems that combining strengths of different methods is the most promising way to account for the multi-faceted nature of the bias.

The methods should also catch up with the changes in measurement trends, e.g. the advent of computerised testing in ILSAs (e.g. PISA is completely computerised since the 2015 edition) makes RT and log file analyses (backtracking, editing, mouse moves) much more accessible and hence much more promising for future studies in the response bias field. Moreover, also the appearance of more advanced psychometric techniques should be seen as a chance to refine self-report measures.

Especially the methods based on the IRT family contributed greatly to response bias research, however, the accomplishments are so far greater in the cognitive tests field (e.g. Stager et al., 2020). Both paradata and internal remedy methods are especially promising in the context of ILSAs/NLSAs as their use does not pose any effort on the side of respondent which is a requirement of paramount importance in this context. Moreover, these methods seem to be less burdened by cultural factors which is crucially important for international projects[58]. The specificity of large-scale assessments also readily balance these methods' drawbacks, e.g. need of large samples, large response vectors and sufficient analytical background. Efficient use of SDR control methods requires also enhancing data patterns interpretations which need to be based on advances in comprehension of the SDR nomological network. Psychophysiological methods can be a great help in future studies on SDR mechanisms.

Nonetheless, in contrast to the many above-presented methods the overclaiming technique seems a promising tool to account for response biases in self-report research. However, the method still needs extensive validation studies in order to establish its practical values as well as to identify processes leading to certain OCT scores, which is indispensable to verify its validity. The rest of the dissertation is dedicated to such a validation study of this method on the example of the PISA 2012 data.

4.6 Chapter summary

Self-report is an easy-to-use and very cost-efficient research method. Both of these advantages explain its immense popularity in a wide array of research disciplines and measurement contexts. However, apart from great virtues the method is also subjected to a heavy critique regarding its validity. Some of the researchers even claim that self-reports collect "Platonic ideal" type of data, as they are often used to gauge unobservable phenomena as opinions, attitudes or judgments of, allegedly, dubious validity.

Despite certain caveats self-report validity has been supported by recent meta-analyses and meta-syntheses. Of course, the method is not free from methodological problems, of which systematic measurement errors are among the most important. One of the prominent sources of such errors are response biases commented in more detail in this work, but apart from them the method has certain moderators of response validity. Most important of them are gender, age, cognitive abilities and participants' interest and motivation to participate in a survey responding. Also item and scale properties play key role in responses validity as more specific and objective items yield more valid scores.

Self-reports are also widely used in the assessment of skills, including academic or school-related abilities. Validity of such scales is well evidenced by a firm body of criterion-validity studies. However, this subfield is not free from its specific problems as subjects often tend to misestimate the level of their skills at question (e.g. math abilities). Interestingly, hard tasks tend to be over-rated, while easy tasks are more often under-rated. Overall, there is a certain tilt toward overly positive self-reports, some studies evidence that only 50% of self-reports is accurate with around 30-35% of self-assessments that are overclaimed. Interestingly, the group of underclaimers is also frequently identified in the field which to date was not adequately addressed by theoretical explanations.

Most importantly, still no consensus has been forged on how to account for this positivity bias in self-report of skills. Many methods were devised and tested but none of them reached status of a "golden", commonly accepted standard. The subchapter 4.5 reviews these methods and contains proposition of

[58] Especially that nowadays such projects entail not only ILSAs but also multi-lab projects, including multi-lab replication research.

a refined classification of methods, divided into preventive and remedial procedures. The conducted review evidenced that there is indeed no method that is adequately verified to be a valid and flexible curative for positivity bias. It is advised to combine and converge many procedures in order to account for response biases of which self-assessment of skills is definitely not free. Due to its easiness and efficiency of use, as well as sound theoretical justification, overclaiming technique is seen as a good candidate for an “all-rounder”-like method available to validly account for positivity bias and other response biases in a wide variety of measurement contexts.

EMPIRICAL PART: OVERCLAIMING TECHNIQUE AS A MEASURE OF POSITIVITY BIAS

Chapter 5- OVERCLAIMING TECHNIQUE AS A MEASURE OF POSITIVITY BIAS- HYPOTHESES DRAWING

5.1 Definitions and similar terms

The above-presented review of the methods conceived to prevent positivity biases or to control the spurious variance they cause in self-reports showed that none of the methods available is a perfect solution for the problem. The ideal method should be easy to use, cost- and time-effective as well as flexible in order to be usable in every research context and every mode of data collection. Specifically, such method should be feasible not only in small research projects, but also in large-scale measurements like ILSAs (e.g. PISA, TIMSS, PIAAC) or major international surveys (e.g. WVS, ESS, etc.).

The overclaiming technique (OCT) could be seen as such a response to methodological problems caused by positivity bias and alike processes. The method was first used by Phillips and Clancy (1972) and was based on experiences from linguistic word recognition tasks (see Zimmerman, Broder, Shaugnessy & Underwood, 1977 for description). OCT is based on presenting to respondents truly existent items ("reals") mixed with items that do not exist ("foils"). The respondents' task is to assess their familiarity or knowledge with the items presented. Claiming knowledge or familiarity with non-existent items is believed to be an index of positivity bias and can be further used in statistical analyses. Hence, overclaiming was defined as *asserting knowledge of a concept that does not exist* (Phillips & Clancy, 1972) or *claiming knowledge about non-existent items* (Paulhus et al., 2003).

On the basis of responses to reals and foils different scores (indices) could be calculated. Paulhus and co-workers advocated (2003) use of Signal Detection Theory (SDT) framework in order to do so, however, other methods were also widely used (e.g. Bing et al., 2011; Muller, 2019; Vonkova et al., 2018). Most importantly, there are two main indices that can be constructed from any OCT: a) accuracy index and b) bias index. Any accuracy index, no matter the method used to calculate it, entails information about how well a given participant discriminates between reals and foils[59]. Hence, this index encodes not only how highly a responder claims her familiarity with reals but also how well does she refrain from claiming familiarity with foils. The most recommended accuracy index based on SDT calculations is *d'* (Paulhus & Petrusic, 2010). Bias indices convey information on how eager a respondent is to claim familiarity with any type of stimuli. This kind of indices are often named "yes-rate" indexes or "location criterion". The most often encountered bias index based on SDT is *c* (Paulhus & Petrusic, 2010). Detailed information on the use of indices will be provided in the methods section below.

Some researchers also use the term "overclaiming" interchangeably with other notions like e.g. "overconfidence", "overestimation" or "bullshitting" (e.g. Jerrim et al., 2019). Frankfurt (2005) defined "bullshitting" as *claim knowledge or expertise in an area where they actually have little experience or skill,* thus situating it close to the definitions of overclaiming proposed by Phillips and Clancy (1972) and Paulhus and co-workers (2003). However, these terms should not be confused with overclaiming as there are measured by different paradigms and have their own definitions as overconfidence is the difference between confidence of correctness and real accuracy (see Pallier et al., 2002) and bullshitting is defined more precisely as *deceptive misrepresentation, short of lying* (Black, 1983) or *pseudoprofound communication, that attempts to impress rather than inform, to be engaging, rather than instructive* (Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2015). Thus, main difference between

[59] These indices are also often named "sensitivity" or "discriminability" indices.

bullshitting and overclaiming is in the complexity and intentionality being much greater in the former than in the latter. Overconfidence has been verified as being only weakly related to overclaiming, defined as it is done here, by Bensch and collaborators (2017). Still it is unknown, how exactly overclaiming is related to tendencies to bullshit or fake news gullibility but some evidence exists that they are somewhat related (Pennycook & Rand, 2020). Conceptually, overclaiming is related to yet another similar term, namely yielding pseudo-opinions as defined by Bishop and co-workers (1986; see also Herr, Sherman & Fazio, 1983). **The definitions presented above situate overclaiming, at least in the context of survey methodology, as a term very close to positivity bias (Bensch et al., 2017). Overclaiming technique (OCT) is, therefore, a measure of the fact and degree of overclaiming in self-report data.** The small differences between the existing OCT versions (Ackerman & Ellingsen, 2014; Goecke et al., 2020; Hargittai, 2005; Ziegler et al., 2013) are ignored and not commented here, as not relevant to the main topic of the dissertation.

Having this terminological clarifications in mind it has to be said that OCT meets the requirements of research efficiency posed before a method targeted at gauging the positivity bias in self-report data. OCT is easy to use, not very time consuming, context- and mode-flexible and have an appealing clear operationalisation of scores. After all, what can be more diagnostic of creating too positive, too desirable characteristic of oneself than claiming familiarity with non-existent items or fictitious skills?

However, this first-glance appeal of OCT has to be verified empirically. To date the results of its usefulness are mixed as many research results point to the lack of convergence validity between the OCT scores and other typical measures of positivity bias (e.g. Ludeke & Makransky, 2016; Muller, 2019). Moreover, there is evidence that OCT scores fail to act as a suppressor or moderator in the relation of the self-report scale scores and a given criterion (e.g. Yuan et al., 2015). However, other research results point to the usefulness of OCT, in example to control for spurious variance in cross-country research (Kyllonen & Bertling, 2013; Vonkova et al., 2018) or to act as suppressor or moderator in criterion-related validity studies (e.g. Anderson et al., 1984; Pokropek, 2014; Yang et al., 2019).

Nevertheless, there is a large call in the field for more validity studies of the method, especially criterion-related validity research (Bing et al., 2011; Ludeke & Makransky, 2016; Paulhus, 2011) and construct validity investigation, in order to continue the attempts to get to know the mechanisms leading to OCT scores and to draw the full nomological network of the method (Bensch et al., 2017; Tonković et al., 2011). The debate on overclaiming mechanisms is still open and inconclusive (e.g. Muller, 2019) thus this research aims to contribute to it by conducting a comprehensive review of the evidence gathered to date and also by providing new analyses based on a rich secondary dataset (PISA 2012 for Poland).

The subsequent parts of this chapter will provide an additional review of research results concerning OCT along with a methodological commentary regarding OCT scoring and applying in analysis. Moreover, research questions will be presented and their justification in the light of the literature will be offered.

5.2 Research review and research questions derivation

5.2.1 Overclaiming scores as a suppressor of spurious variance

First of all an attempt will be made to bring further evidence of a potential suppressor effect of the OCT scores on the relation between self-report scale and related criterion. In this research the math familiarity scale from the PISA 2012 dataset will be used as a self-report scale and PISA test math score will be used as an objective criterion. The assumption is that self-report should correlate positively with the criterion at least in the range of 0.30-0.40 of the standardised correlation coefficient (zero-

order correlation case) as evidenced by reviews of self-report validities (e.g. Ackerman et al., 2002; Mabe & West, 1982; Zell & Krizan, 2014). This research question also aims to verify one of the key effects of positivity bias- suppression (Ganster et al., 1983) and verify mixed evidence present to date in the literature where there is a large evidence base for suppressor effect of OCT scores on criterion-related validity of self-report scales (He & van de Vijver, 2016; Pokropek, 2014; Vonkova et al., 2018; Yang et al., 2019) but there are also some contradictory results that need verification (Bing et al., 2011; Yuan et al., 2015). Moreover, some of the positive evidence was very preliminary and cannot be treated as a full proof. Hence, in light of the theory it is assumed here that OCT scores will act as a classical suppressor (Paulhus, Robins, Trzesniewski & Tracy, 2004; Tzelgov & Henik, 1991) for the relation between math familiarity self-report scores and cognitive (math) test scores, resulting in elevated relation between the self-report and the math test and negative relation between the OCT score and the math test after OCT scores will be introduced to the regression equation. As warranted e.g. by Anderson's and colleagues research (1984) adding OCT scores to the equation should bring exactly these results: negative correlation between inflated score and objective test and boost of R^2 as an indicator of increment in validity of the self-report. Thus:

Hypothesis 1: Overclaiming scores will act as a classical suppressor in the relation between math familiarity self-report scale and objective criterion- cognitive math test.

Positive verification of these hypothesis will provide evidence that OCT can be used to enhance the validity of self-report measures in low-stakes assessments.

Moreover, the research proposed in the subsequent parts of this work is aimed to enlarge our understanding of the mechanisms of OCT. Initial propositions were very optimistic indicating that OCT can measure positivity bias (e.g. Paulhus et al., 2003) but further research brought a lot of contradicting results to this (overoptimistic?) statement (Hulur, Wilhelm & Schipolowski, 2011). For example, Dunlop and colleagues (2017) summarised the research and proposed that overclaiming can result from four different mechanisms: a) self-enhanced self-presentation, b) faking/impression management/lying, c) memory bias, d) careless responding. This summary is a roadmap for testing various hypothesis regarding overclaiming mechanisms.

5.2.2. Overclaiming as a result of memory bias

Huber (2017) specified that the alleged "memory bias" can be due to two different processes: a) false recognition or b) false recollection[60], whereas Calsyn, Kelemen, Jones and Winter (2001) hypothesised that OCT scores may be related to general or specific knowledge of a respondent. This idea was further supported by other researchers who investigated the relation between general and domain knowledge, as well as IQ and openness to experience personality trait[61] and OCT (e.g. Dunlop et al., 2017). All these hypothesis (memory bias, knowledge, openness to experience) have a common core as in the case of the relation between respondent's knowledge (openness to experience) and overclaiming it is assumed that the larger the knowledge, the bigger the tendency to overclaim due to

[60] Recognition and recall are two distinct types of memory retrieval. Recognition is simpler and easier as it entails only recognising which of the presented (incoming) stimuli is known (memorised) and which is unknown (new, not memorised). Recall is also a process of memory retrieval but it is defined as a self-standing process based on retrieving information from memory without any additional cues nor stimuli (Yonelinas, 2002). Recognition can be further divided into "feeling of knowing" (recollection) and "feeling of familiarity" (familiarity) (Mandler, 1980).

[61] Openness to experience is one of the main personality traits (one of the five in the Big5 model) and is often related to IQ and knowledge but also to creativity and non-traditional, unorthodox thinking (e.g. Moutafi, Furnham & Crump, 2006).

larger probability of false recognition of foils due to interference between foils and notions present in the memory system.

Another feasible explanation within the larger "memory bias" framework is that more creative, open to experience people have simply more associations with foils, e.g. due to having richer and/or more disinhibited semantical networks. Hence, creativity and originality in thinking can propel those participants to make sense of foils and claim their familiarity (Muller, 2019). It is worthy to remember that one of the three facets of openness[62], unconventionality, can predict non-traditional thoughts and life decisions (Schwaba, Robins, Grijalva & Bleidorn, 2019). High openness to experience also predicts high verbal fluency, so participants high on unconventionality trait can have more atypical associations with a given foil than other respondents and, consequently, larger chance of claiming familiarity with it which would result in achieving a higher bias score in OCT.

In general, OCT does not display consistent relations with personality (Dunlop et al., 2019). Kam and colleagues (2015) showed little overlap between OCT and personality traits, similarly Barber, Barnes and Carlson (2013) found no correlation between OCT bias index and neuroticism (HEXACO-measured) and survey enjoyment. Much alike results were also achieved by other researchers (e.g. Dunlop et al., 2017; Lee, 2016). Williams, Paulhus and Nathanson (2002) reported no relation between OCQ and self-report measures of perfectionism, clarity of self-concept and optimism. Dunlop and collaborators (2019) reported small to moderate correlations between OCT indices and the HEXACO-model personality traits: bias correlated with extraversion, agreeableness, conscientiousness and openness (*r*s in the range of 0.15-0.25), whereas accuracy correlated only with extraversion and conscientiousness (*r*~0.15). Both OCT indices correlated negatively with emotionality (*r*~ -0.15).

However, one of the main personality traits, openness to experience (and similar measures), is regularly linked to have relations with overclaiming. In example, Ziegler and co-workers (2013) found a positive correlation between bias index and openness (*r*=0.30) and risk-taking traits. Similar correlations between bias index and openness were also found by Tonković and colleagues (2011; *r* ~0.25). On the other hand, Dunlop and co-workers (2017), Ludeke and Makransky (2016) and Swami, Papanicolau & Furnham (2011) found a relation between openness and accuracy index (*r*=0.26), but no relation between openness and OCT bias[63].

Interesting results were presented by Ludeke and Makransky (2016) where bias was also a significant predictor of openness exaggeration, but only in high prestigious domains of OCT (e.g. jazz music, but not rap music, philosophers, but not cloth retailers, etc.). Hence, the content of the scales can influence OCT scores and validity. Items' desirability and instrumentality to make a certain impression is a probable moderator of the OCT-openness relation, along with measurement context (Dunlop et al., 2019). Correlation between openness to experience and accuracy index is probably related to accumulated general knowledge, however, both direct and indirect, knowledge-mediated, paths relating openness to OCT exist. It is important to note, however, that not only greater knowledge could be responsible for the openness-overclaiming correlation, it may be also due to self-enhancement tendency inflating scores in both measures (Dunlop et al., 2017).

For example, Atir, Rosenzweig and Dunning (2015) presented results where overclaiming foils was positively related to self-reported knowledge in a given domain. This result could point to the possible relation between overclaiming and knowledge in line of the memory bias hypothesis. However, Atir

[62] According to one of the models these are: intellectual interests, aesthetic interests and unconventionality, cf. Christensen, Cotter & Silvia (2019).

[63] Interestingly, in two of these studies openness was self-rated, whereas in the third one it was measured by peer ratings.

and colleagues attributed OCT bias to lack of metacognitive skills as in the case of Kruger-Dunning effect (1999), especially that in another set of studies conducted by these researchers participants competent in a certain knowledge domain overclaimed familiarity with foils less than incompetent participants did (Atir, Rosenzweig & Dunning, in preparation). These results would point that the positive relation between OCT bias score and self-rated knowledge is a spurious one, as objectively measured knowledge is related negatively to bias.

Other research evaluated relations between OCT indices and various knowledge and intelligence measures. Swami and colleagues (2011) obtained positive relation between both OCT accuracy and bias index and self-rated intelligence and psychiatric knowledge ($r \sim 0.30$). Pesta and Poznanski (2009) showed that OCT accuracy correlated with an IQ test (r=0.64), exam scores (r~0.40-0.50) and GPA grades (r=0.30). Overclaiming also offered some small incremental validity in predicting MBA exam scores. In the study conducted by Bertsch and Pesta (2009) both OCT indices were related to IQ-accuracy (d') correlated around 0.40, whereas bias (c-measured) was non-significant or negative, however, different bias index, mean foils index, was related negatively to IQ (r=-0.16). Positive relation between OCT accuracy and intelligence were also presented by Bensch and co-workers (2017) (r=0.38) for crystallised intelligence and also Ziegler and colleagues (2013) who did not obtain relation between VOC-T bias and fluid intelligence, but found a small correlation between accuracy and intelligence tests (r~ 0.15). Moreover, Deffler, Leary and Hoyle (2016) found a small positive relation between educational level and accuracy (d'), but no relation between educational level and bias (c-measured). Musch and collaborators (2012) found that OCT accuracy (d') correlated with vocabulary test (r=0.40), whereas Calsyn and others (2001) found that the number of real agencies named in a free recall correlated with agencies overclaiming (r=0.12) and true agency recognition test correlated 0.60 with OCT accuracy. These findings are to a large extent similar to the finding of Stanovich and Cunningham (1992) and West and Stanovich (1991) who obtained significant correlations between their version of OCT and various vocabulary, knowledge and intelligence tests (PPTV[64], Nelson-Denny vocabulary knowledge test and Raven matrices) as well as SAT[65] (r=0.50) and GPA (r=0.35) scores. Very interesting results were showed on the basis of the PISA 2012 dataset by Kyllonen and Bertling (2013) who reported positive relation between OCT accuracy and math test scores, both on individual- and country-level. However, it is worthy to note, that only after the correction for overclaiming[66] this latter correlation achieved significance (from 0.17 to 0.58) and that the correction did not affect the individual-level correlation (0.45 unadjusted, 0.44 adjusted).

In other studies both accuracy and bias indices were related to IQ measures (accuracy- r~0.50, bias- r~0.20; Paulhus et al., 2003; Paulhus & Harms, 2004). Paulhus and Harms (2004) interpreted that more intelligent participants had higher accuracy due to their greater knowledge and cognitive ability and that they had higher bias because of "overgeneralised cognitive confidence". This line of argumentation was supported by Kuncel and co-authors (2012) who proposed that OCT was related to associative memory hence more knowledgeable people should yield higher accuracy and higher bias indices due to overgeneralisation of the associations from their long-term memory[67]. The results presented by Bensch and co-workers (2017) where IQ (crystallised) correlated not only with OCT but

[64] Peabody Picture Vocabulary Test- a popular test for assessing richness of vocabulary.

[65] Scholastic Assessment Test- a popular standardized test for college admissions in the USA.

[66] To this end a following measure was calculated: mean rating for reals minus mean rating for foils (cf. Atir et al., 2015).

[67] "(…) "memory is associative and people with more knowledge are more likely to have many associations. One likely consequence is that knowledgeable people are more likely to have legitimate but inaccurate recognition for foil items" (p. 107)." (Kuncel et al., 2012)

also with a overconfidence measure (*r*=0.46) may point to such metacognitive explanations of OCT scores.

However, this "associative" theory was questioned by other evidence, e.g. Muller (2019) points that the evidence for OCT-intelligence relation is limited, whereas in the work by Hulur *et alii* (2011) overclaiming was not related to IQ (both fluid and crystallised) which pushed the authors to a conclusion that indeed foils functioned "as intended" as they are not measure of e.g. similar sounding overgeneralisation, resulting in a memory bias of false recognition but a measure of positivity bias. Similarly, Franzen and Mader (2019) obtained a negative relation between IQ test and overclaiming tendency. This result corresponds to the result obtained by Bishop and colleagues (1986) where the more a person knew about a subject, the less likely she was to give an opinion about a fictitious political issue. Likewise, Paulhus and Dubois (2014) found that OCT accuracy is positively related to final school grades (*r*=0.37), whereas OCT bias is negatively related to them (*r*=-0.18)[68], namely that again participants with higher knowledge overclaimed less.

On the other hand, null results also appeared in the field where no relation between OCT bias and objective or self-report measures of intelligence, ability, etc. was found: Bing et al. (2011) reported no relation between GPA, ACT[69] and OCT bias, Musch et al. (2012) found that bias did not correlate with discrepancy between self-report and objective ability test, as indicated by the residuals between self-ratings and Abitur[70] results, in other words OCT bias failed to moderate this relation.

Interestingly, there is also certain knowledge lacuna regarding whether self-perceived and objective knowledge have similar or different relation to OCT indices. Atir and colleagues (2015) found that both types of measures (self-reported and objectively measured) predict a positive relation between knowledge and bias (more knowledge, more bias). Moreover, both types of measures do it to a similar degree in case of effect size, there is also a preliminary evidence than their influence is at least in part independent from each other as both regression coefficients were significant when entered into a regression equation. However, Van Prooijen and Krouwel (2019) obtained relation between OCQ bias to self-perceived (0.11), but not actual (test-measured) political knowledge. As only a handful of studies included both types of measures it is impossible to inform this discussion further. However, the PISA dataset gives opportunity to compare relation between both self-reported and objective assessments of within-domain knowledge with OCT indices.

The above review leaves us with two divergent predictions: one that OCT scores are mainly driven by overgeneralisations of existent knowledge leading to false recognitions (memory bias hypothesis), other that these scores depend on metacognitive abilities to monitor and control one's knowledge (metacognitive hypothesis). The former predicts that participants scoring higher on knowledge, ability or intelligence measures would have higher scores on both accuracy and bias OCT indices in comparison to participants scoring lower on these measures (e.g. Atir et al., 2015; Kuncel et al., 2012; Paulhus & Harms, 2004). According to e.g. Dunlop and colleagues (2017) and Paulhus (2011) it can be assumed that more knowledgeable participants accumulated more knowledge, hence they have more illusions of alluring familiarity that propels them to claim familiarity with foils. Simply put, they know more so they have more associations and hence more occasions to commit an error of false familiarity. According to many researchers this account explains the relation between years of education, openness and overclaiming tendencies (Dunlop et al., 2017; Muller, 2019).

[68] Paulhus suggested interpretation in the line of self-enhancement theory as a result of "narcissistic self-destructiveness".
[69] American College Testing- a popular standardized test used for college admissions in the USA.
[70] High-school final exam in Germany. The equivalent of Polish or Austrian Matura.

On the other hand, Atir and co-workers (2015) theorised that this pattern of results may also emerge as a result of motivated "confirmation-biased" memory search or as an outcome of activated expectations about one's knowledge. Inflating one's perception of knowledge, e.g. due to experimental manipulation, leads to inflated rate of false alarms (Atir et al., 2015). The exact mechanism is not known, but it may be explained in the light of the Theory of Planned Behaviour (TPB; Ajzen, 1985; 1991). According to this theory perceiving a given situation as easy leads to development of domain-specific self-efficacy that may contribute to overclaiming in a way of a self-fulfilling prophecy: "I know a lot, so I should also know that". This account is additionally corroborated by the fact that indeed relation between knowledge and overclaiming is domain-specific (Atir et al., in preparation)[71].

The "metacognitive abilities" account predicts that more knowledgeable responders should have higher OCT accuracy, but lower OCT bias in comparison to other participants, exactly as in the study of Dunlop and others (2019) where bias (c-measured) was related to providing opinions on bogus and obscure statements in a test of political knowledge while accuracy was related negatively to provide such opinions. This account is based on the findings showing that OCT accuracy index is a valid measure of one's genuine knowledge/ability and that bias is low in highly competent participants due to their supreme abilities to search one's memory or to inhibit the alluring foils (Atir et al., in preparation; Stanovich & Cunningham, 1992; Stanovich & West, 1989).

In line of this account would be also findings relating OCT bias with fake news gullibility and susceptibility to populist political claims (Pennycook & Rand, 2020; Van Prooijen & Krouwel, 2019). This could mean that overclaimers may just have a very low threshold to accept "anything" as a signal (valid news, existing notion, etc.). However, this claim is doubtful in the light of the alleged domain-specificity of OCT (Atir et al., 2015), especially accounting for low cross-domain correlations for OCT indices, e.g. Calsyn and others (2001) found that overclaiming estate agencies names and political issues was only modestly correlated (r=0.27), but overclaiming agencies names and scientific terms was practically not correlated (r=0.08) and Franzen and Mader (2019) reported low correlation between two versions of their OCT task (r=0.20). These findings seem to indicate that there is no "overclaiming trait" that would predict yielding large OCT bias in every domain.

Additionally to the mentioned studies presenting different relations between measures of competence and OCT indices there is also a large body of studies showing null results- no relation between OCT indices and knowledge/intelligence/ability of participants was achieved e.g. by Bing et al., 2011, Hulur et al., 2011, Jerrim et al., 2019 or Musch et al., 2012.

Disentangling between contradictory evidence is not possible without a set of dedicated experimental studies, however, in case of the PISA 2012 database it is possible to test some of the hypotheses that would indirectly provide some initial evidence on the viability of one of the above hypotheses. Namely, a correlation between math ability, certain self-report characteristics and OCT indices (bias and accuracy) will be calculated in order to assess whether (and if, how) general knowledge (as indicated by math test score and self-reports) is related to OCT indices. To this end three self-report scales, self-efficacy, openness to problem-solving and perseverance, were chosen basing on the relation between these traits and knowledge/cognitive ability, as suggested by the literature (Furnham & Chamorro-Premuzic, 2006; Ziegler, Danay, Heene, Asendorpf & Buhner, 2012; Zhang & Ziegler, 2015). In this way, these self-reports serve as additional measures of math ability, widening the scope of analysis. This enables to test following hypotheses:

[71] As it is between self-efficacy and objective knowledge (e.g. Borgonovi & Pokropek, 2019).

Hypothesis 2: Math ability, as measured by the PISA test, will be positively related to OCT accuracy index and negatively to OCT bias index.

Hypothesis 3: Traits related to math ability, self-efficacy, openness and perseverance, will be positively related to OCT accuracy index and negatively to OCT bias index.

Confirming both hypotheses would support the "metacognitive monitoring and control" proposal (e.g. Dunlop et al., 2019) in contrast to the "overgeneralised associations"/"confirmatory bias/memory bias" prediction (e.g. Kuncel et al., 2012; Paulhus, 2011). It would also put in doubt the self-perception/self-efficacy explanations (Atir et al., 2015).

Moreover, the ample PISA 2012 dataset gives an opportunity to compare relation between self-reported and objectively measured knowledge on one side and OCT scores on the other. Under the memory bias account there should be no difference between such relations, however, according to the motivated bias account self-assessments should be distorted by the same processes as OCT, but objective tests should be absent from them. Hence, if the memory bias account is generally correct it could be expected that:

Hypothesis 4: Self-report math ability assessments will be related to OCT indices similarly and to a similar magnitude as objective measures of math ability.

Answering the hypotheses presented above will help to shed light on the theories attributing overclaiming to memory-based mechanisms.

5.2.3 Overclaiming as result of deliberate response manipulating (faking, impression management, lying, etc.)

Overclaiming can be also result of deliberate response manipulation, be it faking, impression management or simple lying. However, research results pointing to such relation are mixed at best.

Bensch and colleagues (2019) provided evidence that OCT is not related to experimentally-induced faking (conscious distortion) and were also surprised that OCT bias correlated negatively with faking good (*r*= -0.15, this was also significant in a regression equation with *beta*= -0.16). However, elimination of just three participants with a maximal overclaiming score resulted in non-significance of this parameter. Similar results were presented by Feeney and Goffin (2015) who showed that OCT did not correlate with self-reported faking tendencies and only weakly correlated with the measure of self-report distortion called Residualised Individual Change Score (RICS)[72].

On the other hand, Dunlop and colleagues (2019) showed that bias index from OCT was correlated with positive self-presentation behaviours during job interview (r=0.23) but accuracy index was not related to them. Moreover, they have also found a relation with other measures of faking- RICS and within-person correlation (WPC)[73]. The results provided by Ludeke and Makransky (2016) and Muller and Moshagen (2019a) pointed to no relation between OCT and faking, but the results presented by Bing and colleagues (2011) demonstrated that OCT might be used as an indicator of faking tendencies. On the other hand, Joseph, Berry and Deshpande (2009) found a negative relation between self-reported frequency of unethical conduct at work and OCT bias (*r*= -0.21; more ethical conduct correlated with more overclaiming), meaning that overclaiming participants had tendencies to inflate

[72] RICS are the residuals obtained by regressing self-report scores from applicant condition (high motivation to fake good) on the same participant's scores from honest condition (low motivation to fake good) (Feeney & Goffin, 2015). The measure is a widely accepted indicator of faking (Burns & Christiansen, 2011).

[73] WPC is simply a correlation coefficient between two assessment conditions, e.g. fake good and honest responding. Low WPC is interpreted as an indicator of high faking (Burns & Christiansen, 2011).

their reported ethical behaviour. Probably the pattern of results is context-dependent and only job-relevant OCQs can serve to measure faking tendencies that should manifest in higher bias indices in faking conditions (Dunlop et al., 2019; Feeney & Goffin, 2015). This line of argumentation was also supported by Paulhus (2011) who suggested that OCT is more sensitive to context and content than previously thought.

Other evidence on the relation between OCT and deliberate response manipulation is also mixed. Paulhus and Harms (2004) found a correlation between bias and peer-ratings of bragging and egotism (r ~ 0.30), whereas Franzen and Mader (2019) found no relation between OCT scores and intrusive questions, regarding e.g. shoplifting, excessive alcohol consumption, etc (similar results in Randall & Fernandes, 1991). Ludeke and Makransky (2016) also showed absent convergent validity evidence for the OCT and faking as it did not correlate with the self- *versus* peer-ratings discrepancies. Finally, no correlation was found between cheating in an experimental paradigm (coin tossing, dictator's game) and OCT indices (Muller, 2019, Study 2; Muller and Moshagen, 2019) nor between OCT scores and cheating in a knowledge test (Steger et al., 2020). This lack of relation to seemingly valid indicators of faking seriously undermines OCT utility and questions its validity to measure impression management.

Experimental evidence against linking overclaiming and impression management was provided by Atir and colleagues (2015) who manipulated participants' knowledge about the presence of foils in overclaiming task. Warnings reduced the overall bias and had no effect on accuracy (as in Paulhus et al., 2003) but did not have any influence on the relation between participants' declared knowledge and OCT bias, thus undermining the probability that overclaiming is due to conscious response manipulation.

Due to lack of needed data in the PISA 2012 database the hypothesis of OCT as a result of conscious response manipulation (faking, cheating) cannot be verified here. Nevertheless, in the light of the research review presented in the above chapters it is unlikely that a large proportion of respondents (if anyone at all) would engage in this behaviour in a low-stakes educational assessment as PISA. Thus, this mechanism will not be commented further in this work.

5.2.4 Overclaiming as result of non-deliberate, motivated response biases (e.g. self-enhancement tendencies)

Having commented on memory biases and impression management as two potential mechanisms of overclaiming it is time to proceed to the two remaining ones: overclaiming as a result of self-enhancing tendencies, namely overclaiming as a result of motivated processes leading to overly positive presentation in self-report, and overclaiming as a result of non-motivated processes, related to response styles (RS) and careless/insufficient effort responding (C/IER).

Let at first deal with the most popular theory that OCT scores are mainly driven by motivated biases in the type of self-enhancement bias and other biases of self-perception resulting in an overly positive image yielded in self-report (positivity bias). The vast majority of studies investigating this relation used measures of Dark Triad personality traits (especially narcissism), self-esteem inventories and SDR scales to indicate their covariance.

Narcissistic participants are predicted to yield overly positive images in a non-deliberate way even in low-stakes measurement contexts (Gabriel, Critelli & Ee, 1994; Gebauer et al., 2012; John & Robins, 1994; Luo et al., 2019; Maaß & Ziegler, 2017; Raskin, 1991), however, the research evidence on relations between this personality trait and OCT is mixed (review: Grosz, Loesch & Back, 2017). Most of the studies found only small correlations (in the range of r~0.10-0.30) between self-report narcissism measures and OCT. In example, Paulhus and Williams (2002) presented results where bias

(c-measured) correlated 0.17 with narcissism (NPI-measured[74]) and did not correlate with Machiavellianism and psychopathy. OCT accuracy index did not correlate with the Dark Triad traits at all. Similarly, Paulhus et al. (2003) found an NPI-OCT bias correlation of 0.35 (very similar results in Paulhus & Harms (2004)). It is interesting that the narcissism measure (NPI) did not correlate with OCT bias in music domain, whereas it did correlate in academic domain (*r*=0.18). It may point to a conclusion that only the domains that are relevant or instrumental for constructing positive self-image are biased by overclaiming, whereas domains perceived as unimportant and/or irrelevant for self-image may not be susceptible to self-favouring distortion (see also Dunlop et al., 2019). Gebauer and others (2012) confirmed this domain-specificity by showing that agentic narcissists overclaimed in agentic domains (e.g. academic knowledge), whereas communal narcissists overclaimed in communal domains (e.g. humanitarian aid, childcare). Other evidence for the relation between OCT and self-enhancement were brought by Luo and others (2019) who obtained correlations in the range of 0.20-0.30 between OCT bias and various narcissism measures (NPI, CNI[75], NGS[76]) and also a 0.35 correlation with better-than-average task. On the other hand, Bensch et al. (2019) reported that OCT bias and narcissism correlated only 0.13, and no correlation was obtained between OCT and Machiavellianism nor psychopathy (all measured with the Short Dark Triad (SD3) self-report scale). Altogether, the evidence for a relation between OCT and narcissism is mixed (Muller, 2019)- some studies found it (Gebauer et al., 2012; Grosz et al., 2017; Paulhus et al., 2003; Paulhus & Harms, 2004; Paulhus & Williams, 2002), whereas others did not (Bensch et al., 2017; Dunlop et al., 2017; Ludeke & Makransky, 2016).

Interesting account was presented by Grosz and others (2017) who reported that the bias index from OCT was related to social-dominance assertive narcissism and to intellectual-ability assertive narcissism, but not to other types of narcissism. What is more, the relation established was very small (*r*=0.12). This finding relates to the results of Anderson et al. (2012) that claimed that overclaiming serves to win or maintain a social position, but not to bolster self-esteem. Mind also that the same aims are commonly related to grandiose narcissism, which is positively related to self-perceived but negatively to actual knowledge (Zajenkowski, Czarna, Szymaniak & Dufner, 2019). This is probably this part of variance that links narcissism, overconfidence, bullshitting and overclaiming. Furthermore, alike OCT and response biases in general, also narcissism was linked to low "cognitive complexity" (Paulhus & John, 1998; Rhodewalt & Morf, 1995), providing yet another common link between the various effects of response biases.

Similarly mixed evidence exists for the relation between self-esteem and OCT. Paulhus reported a 0.30 zero-order correlation between the RSES scores and OCQ bias, which resulted in a 0.22 coefficient in a regression equation with narcissism (NPI-measured) also included (Paulhus et al., 2003). On the other hand, Mesmer-Magnus, Viswesvaran, Deshpande and Joseph (2006) and Kam and others (2015) found a negative correlation between OCT bias index and self-esteem (*r* around -0.16). On the other hand Tracy and others (2009) reported that OCT bias correlated (*r*=0.14) with self-aggrandisement narcissism (NPI-measured) but not with self-esteem (RSES-measured). Hence, evidence for relations between OCT bias and self-esteem are scarce and inconclusive, but pointing to rather small relation, if any at all.

Also no relations were found between OCT bias and other personality traits often linked with SDR. For example, Paulhus and others (2003) reported insignificant correlation between Snyder's self-

[74] NPI= Narcissistic Personality Inventory (Raskin & Hall, 1979; Raskin & Terry, 1988). One of the most commonly used self-report measures of narcissism as a personality trait.
[75] Communal Narcissism Inventory (Gebauer et al., 2012)
[76] Narcissistic Grandiosity Scale (Rosenthal, Hooley & Steshenko, 2007)

monitoring scale and OCQ bias, whereas Sassenrath (2019) found no relation between OCQ bias and the Interpersonal Relations Inventory (questionnaire measuring empathic responses, interpersonal relations, etc.). In the same vein, neither Schoderbek and Deshpande (1996) nor Randall and Fernandes (1991) reported relation between OCT bias and the Ruch and Newstrom's scale measuring ethical conduct. Additionally, Lee (2016), Muller and Moshagen (2019) and Steger et al. (2020) did not glean evidence for any relation between OCT indices and Honesty-Humility dimension from the HEXACO personality inventory.

Despite the hard critique received by SDR scales they were the most often used indicator of relation between OCT scores and overly positive response bias. However, most of the studies found very scarce evidence for the relation between the two measures. For example, Bishop et al (1986) found no relation between answering to fictitious issues and a summed MCSDS score. Ludeke and Makransky (2016) reported no correlations with the SDE subscale from the BIDR. Franzen and Mader (2019) revealed no relation between short versions of the MCSDS and OCT scores, alike Sassenrath (2019) who failed to find any relation between OCQ bias and the BIDR subscales (SDE and IM). Again no correlation between the MCSDS (short version) and OCT was reported by Calsyn et al. (2001). Barber and others (2013) correlated both the MCSDS and the BIDR with OCT and found them unrelated.

Other studies mainly reported low or very low correlations between OCT and various SDR scales: below 0.20 with the IM subscale from the BIDR (Schoderbek & Deshpande, 1996), 0.15 with communal management from the BIDR (Tonković et al., 2011), 0.11-0.18 with the MCSDS and the BIDR (the SDE subscale; Randall & Fernandes, 1991), a very small (r=0.18) relation of the MCSDS and the BIDR (the SDE subscale only) to bias (c-measured) was found by Dunlop and co-workers (2017). Another scale used in this study, SDS-17 (Stoeber, 2001) correlated with bias (c-measured) (r=0.24), but not with accuracy (d'). Mesmer-Magnus et al. (2006) even obtained negative correlation (r= -0.14) between OCT bias and the SDE from the BIDR pointing to a counter-theoretical relation of more overclaiming with less self-deceptive enhancement. Similarly, Muller and Moshagen (2019) also obtained negative relation between OCT and the IM from the BIDR (r= -0.17), which they interpreted as an indication that this subscale measures honesty, rather than proclivity to bias responses (cf. Uziel, 2010; Zettler et al., 2015).

Slightly higher correlations were also noted, e.g. Paulhus et al. (2003) reported no correlation between OCT indices (neither accuracy, nor bias) and the IM & SDD subscales, but found a moderate correlation (r=0.30) between the SDE subscale of the BIDR and OCT bias. Very similar result was obtained by Musch et al. (2012), where the German version of the BIDR was used, and Paulhus and Harms (2004)- in both studies OCT bias (c-measured) correlated only with the SDE, not the IM (r~0.30). Surprisingly, the reversed pattern was discovered by Bensch and co-workers (2017) where OCT bias index correlated with IM (r=0.35), but not SDE (r~0). The authors related this unusual pattern to high correlation between the scales in this particular study, however, did not offer any predictions why this inter-dimensional relation was so high.

The above results show little relation between OCT and typical measures of motivated response biases: Dark Triad personality self-reports, SDR scales, etc. However, it is possible that OCT is driven by self-enhancement tendencies but that the measures used to capture these tendencies simply fail to do that and cannot be treated as valid indicators of positivity bias. Tonković and colleagues (2011) argued that it is the content that drives the relation between the BIDR subscales and OCT, because of the egoistic, agentic content of these subscales, e.g. bragging knowledge or skills. In this vein of interpretation any relation between SDR scales and OCT is only secondary, driven by content- and context-dependent characteristics of the OCT version and the whole study. This interpretation would thus suggest that any relation between OCT scores and SDR scales is only spurious.

It is worthy to gauge if better[77] assessment of self-enhancement tendencies can be achieved by measuring importance and desirability of a given domain for participants. There is evidence that skills that are more valued are more overclaimed (Paulhus et al., 2003; Paulhus & Trapnell, 2008). Furthermore, the research by John and Robins (1993) and Vazire (2010) suggests that accuracy of self-assessment should be higher in unimportant domains rather than important, because in case of important domains it is clouded by motivational factors (cf. Zell & Krizan, 2014). A certain confirmation of this views was provided by Gebauer, Sedikides and Schrade (2017) who conducted a study in which Christians overclaimed more on topics related to their faith (e.g. Bible, saints), less on communal topics, and did not overclaim on agentic topics.

A handy and valid measure of item/domain desirability is self-reported perceived trait desirability (Philips & Clancy, 1972). It is a stronger predictor of overclaiming than other domain characteristics, as presented by Dunlop and others (2019). A broader theoretical and methodological framework related to the trait desirability hypothesis, called ideographically desirable responding (IDR), was proposed and corroborated by Ludeke, Weisberg and DeYoung (2013; see also Sinha & Krueger, 1998). It was showed that IDR was positively related to SDR (BIDR-measured) and self-esteem (RSES-measured), thus confirming the convergent validity of the method. IDR is thought related to the centrality and importance of a given trait for a certain participant, thus more desirable domains/topics should be more overclaimed in OCT (Dunlop et al., 2019). Similarly, less desirable traits should be overclaimed less, in example, MacIntyre, Noels and Clement (1997) showed that respondents with high anxiety towards a school subject tended to underestimate their skills, whereas the opposite was true for those with low anxiety towards this subject.

In the PISA 2012 database there are no commonly used measures of SDR/self-enhancement tendencies, like e.g. SDR scales, Dark Triad inventories or classical self-esteem questionnaires. Similarly, no direct measures of item/domain desirability were used. However, proxy measures could be provided through the use of other scales that should code importance of a given domain for participants. Such scales present in the mentioned dataset are e.g. intrinsic motivation for mathematics, instrumental motivation for mathematics, subjective norms about learning mathematics, intentions about future math education, math anxiety (reversed relation), and various scales related to math and math-related behaviours. If OCT bias scores are in fact driven by self-enhancement tendencies then these scales should correlate positively with them. Moreover, math anxiety and other scales coding negative attitudes towards mathematics should be related negatively to OCT bias. Hence:

Hypothesis 5: Scales related to the perceived importance, effort and joy of learning mathematics will be related positively to OCT bias.

Hypothesis 6: Scales related negatively to the perceived importance, effort and joy, e.g. math anxiety, will be related negatively to OCT bias.

Thus, if the self-enhancement tendencies are responsible for OCT bias scores pupils for whom math is important and central should overclaim more on the PISA 2012 OCT.

There is also evidence that pupils motivated to learn mathematics, but struggling to achieve good results, may be especially prone to develop self-enhancement tendencies. Such students may feel threatened by the discrepancy between their expected and achieved outcomes and resort to self-enhancement in order to save their favourable self-image. This students may also develop attributional theories that would e.g. put the blame of their failures on external criteria or lack of will to succeed

[77] In comparison to, allegedly unsuitable for this task, SDR or narcissism scales.

(the "I could if I wanted to" attitude). Such attitudes should correlate positively with OCT bias and do not correlate or even correlate negatively with OCT accuracy (see Griffith et al., 2006). Thus:

Hypothesis 7: Scales measuring perceptions of success control will be related negatively to OCT bias (less control, more bias) and positively to OCT accuracy (less control, less accuracy).

Moreover, self-enhancement is related not only to "pumping up" virtues but also to diminishing vices. Hence, on the basis of this account, it can be predicted that self-reports of negative phenomena, such as truancy from school, disciplinary problems or negative school attitudes, should be correlated negatively with OCT bias:

Hypothesis 8: Reporting negative school-related phenomena will be related negatively to OCT bias (more negative phenomena reported, less bias).

The above relation was not found by Jerrim and others (2019) for the same PISA data from the Anglo-Saxon countries, here it will be tried to confirm their findings on a different sample and with a slightly different analytical approach.

PISA dataset gives a rather unique opportunity to contrast and compare data on similar phenomena stemming from separate sources of information: school students and school principals. Of course the overlap of the two questionnaires is not great but the answers from both sources can be compared in case of school disciplinary climate. It is predicted that:

Hypothesis 9: The degree of divergence between students' and principals' opinions on school disciplinary climate will be positively related to OCT bias (more divergence, more bias).

Finally, also holding respondents accountable for their answers is related to lower positivity bias in responses (e.g. Lerner & Tetlock, 1999; Sedikides, Herbst, Hardin & Hardin, 2002). However, in the PISA study the participating students cannot be held accountable as it is a low-stakes and anonymous assessment. Nevertheless, information on the school policy in using achievement data in accountability procedures could be used as a proxy for school's attitude towards assessments, PISA included, in general (Vonkova et al., 2018). Is it treated seriously? Or is it treated lightly, or even as an unnecessary nuisance, a whim of country authorities? It is predicted that:

Hypothesis 10: Using assessment data in accountable procedures will correlate negatively with OCT bias.

Hence, the above hypotheses are aimed to test the assumption that OCT scores are driven by motivated tendencies to present oneself in an overly positive light, e.g. as a consequence of self-enhancement or SDR.

5.2.5 Overclaiming as result of response styles and careless/insufficient effort responding

The last mechanism that was suggested to play role in OCT scores emergence is careless/insufficient responding (C/IER). This prediction was voiced by many researchers (e.g. Bing et al., 2011; Bensch et al., 2019; Muller & Moshagen, 2019a) but few took the effort to find an empirical support for such claims.

However, there are several studies that managed to gather some interesting evidence. Barber and others (2013) showed a substantial correlation between OCT bias (measured as averaged foils scores) and bogus and IRIs items (r=0.45 and r=0.51) pointing to a relation between OCT bias and C/IER. The researchers also obtained small correlation (r~0.15) between OCT bias and acute insomnia, sleepiness and self-control depletion (all measured by self-reports). No relation to chronic insomnia was found, though. Here sleepiness (sleep deprivation) was treated as a factor driving C/IER and, consequentially,

related to higher overclaiming due to lowered self-control and cognitive resources[78] (Barber et al., 2013).

Similarly, Ludeke and Makransky (2016) found that OCT bias is related to Mahalanobis distance (r=0.15) and person fit (perfit) measures[79] (r=0.45) and it is unrelated to contra-logical survey errors[80]. The researchers interpreted this rather high correlation between perfit measures and bias as an indicator that overclaiming may be due to C/IER. All C/IER measures (Mahalanobis, perfit, survey errors) were negatively correlated with accuracy (*d'*-measured) further corroborating the hypothesis of importance of C/IER on OCT scores. Calsyn and Winter (1999) used a similar measure of counter-logical survey errors and also did not find any relation between it and OCT bias, but found a small and negative correlation between survey errors and OCT accuracy (r= -.16).

Jerrim et al. (2019) did not calculate any C/IER indices but tried to show an indirect evidence that overclaiming is not related to C/IER, by showing zero correlation between truancy and self-reported test effort with the overclaiming index. As C/IER often causes unrelated measures to correlate with each other (Carden, Camper & Holtzman, 2018) Jerrim took this analysis as an indirect evidence of C/IER relation to OCT scores. Failure to find such correlations enabled Jerrim and co-workers to claim that OCT was not related to C/IER.

Another response bias that may be a potential mechanism of OCT scores are response styles (RS; Jerrim et al., 2019). The evidence on relation between OCT and RS is even scarcer than in case of C/IER. To the best knowledge of the author there is only one study analysing this relation- Dunlop et al. (2019) found that a crude extreme response style (ERS) measure (count of extreme responses) correlated with OCT accuracy (r=0.15) and bias (r=0.23). This would show that OCT scores and RS share some variance but not enough to claim that both are effects of the same mechanism. Another study that used OCT and RS together is Pokropek (2014) but no relations between the two measures were tested there[81]. The study by Pokropek, Khorramdel and von Davier (in preparation) also applied both of these measures together and found a positive relation between RS and OCT bias, which seemed more pronounced in case of ERS than MRS. However, the relation between OCT and RS was not directly measured and was not the focal point of the analyses performed.

Some indirect evidence linking OCT and C/IER and RS stems from twin studies: both RS (Melchers, Plieger, Montag, Reuter, Spinath & Hahn, 2018) and OCT (Luo et al., 2019) were proved to have around 40% of variance due to genetics, few % to shared environment and more than 50% due to non-shared environment (see also similar results on acquiescence response style; Kam, Schermer, Harris & Vernon, 2013). More research is needed to confirm and interpret these similar patterns.

Nevertheless, the existing evidence on the relations between C/IER, RS and OCT is very scarce. From the ample research on C/IER and RS there are many indices and methods that could be used to bring the OCT-C/IER/RS alleged relation to a more thorough test than it has been done to date. In this study

[78] Curiously, these results pointed also to lowered scores on the SDR scales used, showing a lower cognitive control-lower SDR relation, whereas other studies pointed to a reversed relation where low cognitive control was related to higher SDR/self-favouring responding (e.g. Paulhus et al., 1987; Robins & Beer, 2019).

[79] Both Mahalanobis distance and the family of person fit indices are used to detect outliers or to identify aberrant data patterns, namely to identify participants whose data should be treated with extra care due to possible distortion or error (e.g. Conijn, Emons & Sijtsma, 2014). Both group of indices is widely used in the C/IER research (Meade & Craig, 2012; Niessen, Meijer & Tendeiro, 2016).

[80] Indicating that one does not use computer at all and then indicating that one uses Internet for 2 hours a day is an example of this kind of survey error.

[81] Save that the VIF statistic for the regression model analysed in the paper was given, suggesting that there were no problems with collinearity between the predictors (RS and OCT measures). This evidence suggest that correlations between them were not very high (Pokropek, 2014).

a wide and diversified range of C/IER indices will be used to establish their relation with OCT scores. Hence:

Hypothesis 11: Careless/insufficient effort responding indices will be correlated positively to OCT bias and negatively to OCT accuracy indexes.

An additional analysis of the relation between fatigue, C/IER and OCT will capitalise on PISA's missing-by-design structure, where scales are embedded inside of forms and forms are rotated across participants. In different forms specific scales are put in different order, hence some scales are responded to at the beginning of the questionnaire session in one form and on the end of the session in other forms. This design allows to compare OCT and C/IER indices in two conditions differed by the level of fatigue: low, when overclaiming scale was used at the beginning of a form and high, when it was placed near the end of a form. Thus:

Hypothesis 11a: OCT bias will be larger and OCT accuracy smaller in high fatigue condition than in low fatigue condition.

Confirming hypothesis 11 and 11a will provide evidence in favour of the theory of overclaiming as a result of C/IER. Other pattern will yield evidence disconfirming this theory.

In order to examine the relation between RS and OCT the newly developed RS models will be used (Khorramdel & von Davier, 2014; Khorramdel et al., 2019; Pokropek et al., in preparation). Extreme response style (ERS) is predicted to be related positively to both OCT indices (bias and accuracy) (as hinted by Pokropek et al., in preparation). Thus:

Hypothesis 12: Extreme response styles (ERS) will be positively related to OCT accuracy and OCT bias.

The above analyses will inform about the nature of relations between C/IER, RS and OCT using a wide variety of indices and also some newly developed models basing on multidimensional IRT models that offer promise to accurately model the RS variance in self-reports.

5.2.6 Structural validity as an indicator of OCT mechanisms

Another potential source of evidence on OCT's nature is its structural validity. The key indicator is its dimensionality, namely whether the scale is uni-, bi- or multidimensional which could be an indicator of number and kind of processes laying in the foundation of OCT scores. For example, if OCT is driven mainly by self-enhancement or C/IER/RS all items should be correlated with each other and, even if two factors emerge (one for reals, one for foils), they should share a large proportion of their variance (Hulur et al., 2011). However, if OCT is driven by memory biases two relatively unrelated factors should emerge: one for foils, one for reals, basing on the different cognitive mechanisms predicted to be engaged in responding to reals and foils. It may be surprising, but among all the previous studies on OCT only Pokropek (2014) analysed OCT scale's psychometric structure. The results obtained yielded that reals and foils formed separate factors. However, no further analyses were performed as the study had different focus and it was only a preliminary conference paper.

Hence, there is no firm indication regarding OCT scales factorial structure. Some indirect evidence does exist, though it is based on the correlation between OCT bias and accuracy indices. In the study of Dunlop et al. (2019) *c* and *d'* correlated 0.27, but in Ziegler et al. (2013) they were correlated -0.26 (populational sample) and -0.74 (student sample). It is important to note that the sign of the correlation is sometimes reversed. This is due to multiplying the *c* parameter by -1 in order to have higher values indicating more bias (e.g. Dunlop et al., 2019; Muller & Moshagen, 2018). This information has to be taken into consideration when analysing correlations between OCT scores, however it obviously influences only the sign, not the magnitude of coefficients. Other studies also

yielded substantial correlations between the two indices, e.g. in the study by Bertsch and Pesta (2009) the correlation between d' and c amounted to -0.49, in Muller and Moshagen (2018) d' and c correlated -0.46 and in Muller and Moshagen (2019a) this correlation was only -0.20 (c timed by -1). In Mackinnon and Wang (2020) this relation amounted to -0.44 and in Hughes and Beer (2012) it was 0.43 (SDT indexes d' and c were used but no info is present whether they also have changed the sign of *c*).

However, analysing such a correlation conveys certain problems. First of all, there is no firm evidence how *c* and *d*' correlation should be interpreted and whether it is interpretable at all. As the indices are created using the same variables (hit and false alarm rate), a correlation is expected. However, there is no specific correlation size expected nor specified as a coefficient in practically any size can be (theoretically) expected (Dunlop et al., 2017). Moreover, the cross-study and cross-sample stability of this correlation is problematic, e.g. Ziegler et al. (2013) found huge differences in its value from one study to another. As there are no systematic analyses of this topic it is impossible to say what differences between the studies caused this difference in case of the Ziegler et al. (2013) research and in general what measurement characteristics are related to certain values of this coefficient (Ludeke & Makransky, 2016). Additionally, the indices based on small number of items (e.g. only on reals or foils) are inherently in peril of low reliability (Dunlop et al., 2017).

Moreover, there is also a small amount of information regarding correlations between foils and reals themselves. In one of the few studies that shared these values the inter-item correlations were not very high (*r* in the range of 0.28 to 0.50; Franzen & Mader, 2019). On the other hand, Joseph et al. (2009) reported a quite high Cronbach's alpha of 0.78 for an OCQ consisted only of foils on different lay topics. In a similar OCQ version, but this time consisting of foils in 40%, Randall and Fernandes (1991) reported a Cronbach's alpha calculated from both reals and foils amounting to 0.70.

To sum up- it is difficult to pose any specific hypothesis regarding OCT psychometric structure due to scarcity of evidence. However, following Pokropek (2014) who performed initial analyses on the same dataset it is assumed that:

Hypothesis 13: OCT items will form two correlated factors: one for reals and one for foils.

This analysis is largely an exploratory one as lack of informative results prevents forming more specific confirmatory hypotheses. However, the exploration of this topic may bring interesting information on overclaiming mechanisms.

5.2.7 Social norms and overclaiming: school-level analysis

Response biases are also strongly related to the perceived social norms and even to culture in a given society (Kemmelmeier, 2016). Relations between culture, different country-level variables and OCT scores were presented in dedicated publications (e.g. Fell & Koenig, 2016; 2020; Fell et al., 2019; Vonkova et al., 2018). However, relations between smaller groups norms and OCT are largely uncharted. To the best knowledge of the author such relations were analysed only by Jerrim et al. (2019) who observed that only a few percent of OCT variance occurred between schools, indicating that overclaimers are rather "evenly spread across schools". No other school-level variables were analysed in their study. Proportion of OCT indices' variance on individual- and school-level will be also examined in this study.

In this work an effort will be given to explore the possible school-level covariates of overclaiming. There are attempts to explain response biases in the frame of the TPB (Ajzen, 1991) where social norms and subjective beliefs about them are important predictors of behaviour. For a 15-year-old adolescent

school is certainly an important source of social comparisons with teachers' or peers' norms and beliefs constituting an influential point of reference.

A set of variables that could be used in such a school-level analysis was suggested by Vonkova and colleagues (2018) and, among others, included: average level of math ability level at school, average level of math anxiety at school, average level of self-reported math knowledge at school (e.g. math self-efficacy scale, math familiarity scale, etc.), male-female ratio, school type (private *versus* public). To this list a pressure on achievements can be added as a potential threatening factor arousing self-enhancement tendencies (Fell & Koenig, 2016; 2020). It is known that pressure to succeed rises students' academic cheating and faking[82] (Bong, 2008). In the PISA 2012 the pressure variable was measured in the school principals' questionnaire.

Another variable that can be joined to the above list is social status, in PISA measured by a set of self-report items in the student's questionnaire and presented under the form of educational, social and cultural status index (ESCS). This variable was also rarely related to response biases, but recently Jerrim and others (2019) presented an analysis where ESCS covaried with OCT bias, in such a way that higher status predicted higher bias. Nonetheless, no explanation for this pattern was offered in that paper. It is possible, however, that ESCS serves as yet another "pressure" factor- students from high status families may be more prone to parental pressure, upward social comparisons and other factors compelling self-enhancement tendencies. On the basis of this evidence it can be predicted that:

Hypothesis 14: School-level variables indicating high math abilities or high pressure to math-related attainments will correlate positively with OCT bias.

Such pattern of results is suggested on the basis of SDR/self-enhancement theories that predict larger self-enhancement when self-esteem is threatened. In case of highly competent (and competitive) backgrounds, e.g. in schools with high level of math ability and/or high pressure on educational achievements this can be a fact due to high social expectations and social norm to know math very well. Moreover, in such competitive social groups success is greatly admired which may compel certain individuals to find a shortcut to achieve it by overclaiming (Anderman, Griesinger & Westerfield, 1998; Fell & Koenig, 2016; Tett & Simonet, 2011).

Another group of school-level variables that is potentially related to OCT are scales coding possible rule violation and anti-social behaviours, e.g. truancy, grade repetition, disciplinary climate, sense of belonging to school, school-level of overclaiming, etc. Country-level rule violation was identified as a predictor of OCT bias in analyses performed on the PISA 2012 dataset by Fell and colleagues (2019). However, no school-level data was analysed by them. In this work an attempt will be made to confirm these findings also on the school-level using the above-listed variables as proxies for rule violations ((e.g. low school discipline, low sense of belonging, high truancy, high grade repetition, etc.). Hence:

Hypothesis 15: School-level rule violation will be related positively to OCT bias and negatively to OCT accuracy.

Such pattern of results is predicted on the basis of the assumed higher tolerance for questionable moral behaviours in such social groups or lower (insufficient, ineffective) control in them that promote dishonest behaviours such as overclaiming (mild form of academic cheating; Fell & Koenig, 2016; see also Griffith & McDaniel, 2006 on "everybody fakes" attitudes).

[82] See also https://news.stanford.edu/news/2005/february23/cheat-022305.html

The above analyses will offer a comprehensive, though not conclusive, review of different theories on mechanisms of overclaiming. The analyses will comprise both individual- and school-related variables. Some of the analyses planned are largely exploratory, due to scarcity of previous evidence.

5.2.8 Correlates of overclaiming- building nomological network

Apart from testing hypothesis informing on mechanisms of overclaiming another analysis is planned: building nomological network of OCT by analysing its correlates. This enables to inform future theoretical and empirical research on OCT and also offers predictions for future experimental studies.

Socio-demographic variables

First of all, relations between sociodemographic variables and OCT will be examined. One of such variables is gender. The evidence gathered to date offers a very mixed picture.

Gender

Large portion of studies reported no gender differences. Such pattern of results was found in example by: Calsyn & Winter (1999); Feeney & Goffin (2015); Franzen & Mader (2019); Joseph et al. (2009); Mesmer-Magnus et al. (2006); Paulhus & Dubois (2014); Paulhus et al. (2003); Paulhus & Petrusic (2010); Schoderbek & Deshpande (1996); Swami et al. (2011). Dunlop and others (2017) found gender differences only in one out of three studies conducted.

However, in a number of studies using self-enhancement designs some gender differences did appear. In example, Furnham, Zheng and Chamorro-Premuzic (2005) identified that male participants were more prone to better-than-average bias regarding IQ self-assessment than female ones, whereas Bornholt, Goodnow and Cooney (1994) reported that men were more overconfident than women. Also Nuhfer and others (2017) reported that women were better at self-assessment than men, who tended to overestimate their abilities. Moreover, Bishop and co-workers (1986) disclosed that men claimed opinions on fictitious issues more often than women. Similarly, Zhang, Paulhus and Ziegler (2018) found that male students were more prone to scholastic cheating than female students.

In research using OCT specifically such differences also occurred, especially in studies exploring the PISA 2012 dataset. More overclaiming in males was found by Jerrim et al. (2019), Fell et al. (2019) and Yang et al. (2019). Also in other OCT studies this pattern was encountered (e.g. Atir et al., 2015; Dunlop et al., 2017; Ziegler et al., 2013). However, reversed pattern, with women overclaiming more, was also identified (e.g. Calsyn et al., 2001; Philips & Clancy, 1972).

This seemingly discrepant pattern of results can be expected if specific content used to construct OCT items will be taken under consideration. As it was made evident in the SDR research the proposed “types” of this behaviour were driven predominantly by agentic or communal content of SDR scales (Paulhus, 2002). Paulhus and John (1998) stated that women should be more susceptible to communal bias, whereas men should be more prone to agentic bias. There are many other research sources that confirm this relation (e.g. Bem, 1972; Gilligan, 1982; Helgeson, 1994; McGuire, 1968 even contended that women are more susceptible to what is “right” in the society, hence are more prone to communal bias; see also larger concerns for privacy and social approval of women survey respondents as evidenced by Rasinski, Willis, Baldwin & Jobe, 1994). Other factor that should be taken into account before analysing gender differences in OCT is gender-related desirability of a trait/domain measured by a given version of OCT. Paunonen (2015) offers evidence that genders indeed differ in perceived trait-desirability (but see Alicke, 1985, where gender differences in traits desirability and controllability were not found).

Philips and Clancy (1972) followed this line of thinking in interpreting the result obtained in their study were female overclaimed more, although the magnitude of the relation was rather small (Yule's Q=0.15). The authors concluded that their OCT items were related to being "up on new things"- trait that is rated as more desirable by women than men.

The content used in the PISA 2012 OCT involves academic skills and knowledge which is definitely a type of agentic content (Paulhus & Trapnell, 2008). Hence, it should predict higher overclaiming in male participants. However, to additionally inform the predictions it is warranted to quickly analyse the social image of math in Poland with a special focus on any gender-related differences. Baczko-Dombi (2017) in her analyses pointed out that math is perceived as an important but also difficult school subject. Moreover, many school pupils and students do not believe that they can succeed in it. As evidenced by ILSAs' results, Poland is one of the countries with a very equal inter-gender level of math proficiency, e.g. only in PISA 2015 boys scored better than girls by a significant margin of eleven PISA points. In other PISA editions there were no significant differences between the genders in Poland (Sitek, 2019). The gender difference is also very small in case of Polish standardised school exams but gets large and in favour of male students in case of the basic-level Matura exam (final high school exam) (Grudniewska & Kondratek, 2012; Zawistowska, 2013; 2017). Female students also experience higher math anxiety than their male counterparts[83] (Cipora, Willems, Szwarc & Nuerk, 2018; Henschel & Roick, 2017). Female students are also less likely to take extended-level Matura exams in math, even holding constant their cognitive skills (Zawistowska & Sadowski, 2019). These results incline to formulate a conclusion that in Poland, at least from a certain stage of education, math is identified as a "boy thing", hence social norms predict that boys will be characterised by higher math abilities in comparison to girls. Thus, it is hypothesised that:

Hypothesis 16: Boys will overclaim more than girls in PISA 2012 OCT.

Socio-economic status

Other socio-demographic variables were studied in the context of OCT even rarer than gender. Some evidence is available in case of socio-economic status: Jerrim and others (2019) found that students of higher status (as measured by the PISA ESCS index) overclaimed more than their peers. However, most of the difference driven by ESCS was caused by one item ("proper number"), inducing hypotheses that some of the foils may provoke larger overclaiming in certain groups of students. Obviously, no systematic analyses are possible in this matter on the PISA 2012 OCT with only three foils available. However, the status differences were not confirmed by Calsyn et al. (2001) who found that self-reported income was not related to overclaiming and only small and negative relation between educational level and overclaiming was obtained.

Jerrim et al. (2019) found also more overclaiming in immigrant groups in Anglo-Saxon countries (except the USA). However, similarly as in the case of differences in OCT bias regarding the ESCS, the are no readily available explanations for these results. However, due to the relation between ESCS and abilities (Marks & Pokropek, 2019; OECD, 2014a) it can be hypothesised that these differences have similar roots as differences observed between high- and low-skilled participants, if interpreted in the vein of the motivated bias framework (memory bias hypothesis would predict the reversed pattern of relation). In this work both hypotheses will be tested, moreover, an analysis will be performed to check whether ESCS and gender differences are independent from ability level:

Hypothesis 17: Economical, social and cultural status (ESCS) will correlate positively with OCT accuracy and negatively with OCT bias (higher status, lower bias).

[83] This effect holds even when controlled for math proficiency (Henschel & Roick, 2017).

Hypothesis 18: Gender and ESCS relation with OCT indices will be independent from math ability.

Jerrim and others (2019) proposed also interesting analysis of OCT items, namely testing differential item functioning[84] (DIF) (Holland & Thayer, 1986; Swaminathan & Rogers, 1990), regarding ESCS. In this work these analyses will be replicated in the Polish sample and also expanded on other variables, e.g. gender.

Age

Ludeke and Makransky (2016) found no relation between overclaiming and age (group was restricted to 15-30-years-olds only). Similar age groups were also used by e.g. Clariana and others (2016) or Mesmer-Magnus and co-workers (2006), also to yield no age-related differences in OCT bias and only small differences in OCT accuracy in the former study. These differences could be plausibly associated to the knowledge of participants that differed according to the educational level (see also Paulhus, 2011). It is important to notice that most of the studies in the field is based on college or high school student samples so age range in the sample is mostly very limited. Probably the only study that included a wide range of age groups (30-70+) in the sample was Calsyn and Winter (1999) who reported a small negative correlation between age and OCT bias (r= -0.11). Thus, there is only a very limited evidence of age-related differences in OCT. However, this topic is potentially interesting, as age is known to covary with OCT correlates, e.g. narcissism, that is known to change with age (Foster, Campbell & Twenge, 2003). Hence, future research is to verify role of age as a possible covariate and moderator of OCT. In case of PISA 2012 no age differences are predicted as the sample is drawn from a cohort of 15-years-olds and only a small fraction of students in the sample is of marginally different age.

5.2.9 Individual differentiation of overclaiming

Analysing differences between conditions or social groups is very informing on the nature of processes, overclaiming not being an exception. However, also a slightly different point of view is possible, namely exploring latent patterns and groups that are not coded by any observable variables but can be discerned in statistical analyses of latent classes and profiles.

The reasonableness of this approach to analyse OCT is corroborated by results showing that there are important inter-individual differences in response biases that are not captured by standard grouping variables. Bishop and others (1980) and Schuman and Presser (1981) identified that only about 30% of participants give opinion on non-existent issue, the rest does not embrace them at all and hence does not contribute to the spurious variance. Similar percentage of 30% of overclaimers was identified in the research by Randall and Fernandes (1991), though other studies yielded lower values, e.g. 25% in Paulhus and others (2003) and only 6% in Ludeke and Makransky (2016). The proportion of overclaimers seems to depend on foils' desirability and instrumentality[85] and probably other factors which are, as yet, unverified (Dunlop et al., 2019).

Such latent differences between participants pose an additional difficulty the analysis of OCT scores: some respondents can overclaim a lot, while others may not overclaim at all or even underclaim. The effects of under- and overclaimers may cancel out and be skipped in standard OCT analyses, suggesting a null effect, whereas in fact a lot of patterns could be observed if other data analysis techniques had been used. Interesting light on this topic can be shed by using mixture models technique which unable

[84] DIF is identified when in case of respondents on the same level of abilities, but belonging to two different groups (e.g. regarding gender, type of school, etc.), the test responses have different distribution (e.g. Kondratek, Skórska & Świst, 2015).

[85] Instrumentality, utility towards creating a given picture (Ziegler & Kemper, 2013).

to discern latent groups of respondents (different patterns of OCT responses). These techniques revealed differently responding groups in RS research (Khorramdel et al., 2019; Ziegler & Kemper, 2013), in SDR analyses (Leite & Cooper, 2010; Levin & Zickar, 2002), similar techniques were also used in case of C/IER (Meade & Craig, 2012).

In this work a latent class model analysis is planned in order to identify different overclaiming groups among participants and to further examine the characteristics of these groups, e.g. through analyses of their correlations with other variables relevant for overclaiming. Moreover, an analysis of covariates predicting each group membership probability is planned.

As this is a largely exploratory analysis only a rather non-specific hypothesis can be formed, as there is only one study informing on potential latent groups in OCT (Yang et al., 2019). However, the analysis in this work will be performed in order to compare and contrast the results with the study by Yang et al. (2019) who obtained three latent classes using the data from the US sample of the PISA 2012. They interpreted the obtained classes as: a) accurate self-report and high math abilities, b) accurate self-report and low math abilities and c) overly positive self-report and average math abilities. The classes accounted for about 60%, 20% and 20% of the whole sample respectively. Hence, it is supposed that:

Hypothesis 19: Latent class analysis (LCA) models will reveal different subgroups of responders, including a non-overclaiming group, an overclaiming group and an underclaiming group.

Differences of means of relevant variables will be calculated between the emerged latent groups will in order to understand their characteristics better.

The analyses planned in this part will be presented in two subsequent chapters where details of the methods used and all the results will be described. Next chapter, Chapter 6, covers basic methodological information that is common for every analysis presented, e.g. database characteristics. Precise technical details related to a given model or hypothesis are given along with the results presentation in Chapter 7. Such organisation was chosen in order to ease processing of information.

5.3 Chapter summary

The thorough review of the overclaiming research to date led to the formulation of 19 hypotheses that were tested in a series of subsequent analyses. The hypotheses are summarised with their aims in the table below:

Hypothesis	Aim
Hypothesis 1: suppression model	establish whether measurement of OCT bias can lead to increase in self-report predictive validity
Hypothesis 2: objective domain ability and overclaiming	test memory bias hypothesis as a probable mechanism leading to overclaiming
Hypothesis 3 & 4: subjective domain ability and overclaiming	test memory bias hypothesis as a probable mechanism leading to overclaiming
Hypothesis 5 & 6: domain desirability and overclaiming	test motivated (positivity) bias hypothesis as a probable mechanism leading to overclaiming
Hypothesis 7: locus of control and overclaiming	test motivated (positivity) bias hypothesis as a probable mechanism leading to overclaiming
Hypothesis 8 & 9: withholding negative information and overclaiming	test motivated (positivity) bias hypothesis as a probable mechanism leading to overclaiming
Hypothesis 10: school pressure on domain achievement and overclaiming	test motivated (positivity) bias hypothesis as a probable mechanism leading to overclaiming; gauge school-level covariates of overclaiming
Hypothesis 11: careless responding, respondents' fatigue and overclaiming	test response style/careless responding hypothesis as a probable mechanism leading to overclaiming
Hypothesis 12: response styles and overclaiming	test response style/careless responding hypothesis as a probable mechanism leading to overclaiming
Hypothesis 13: latent structure of overclaiming scale	assess latent structure of the PISA 2012 OCT
Hypothesis 14: school-level social pressure on domain achievements and overclaiming	test motivated (positivity) bias hypothesis as a probable mechanism leading to overclaiming; gauge school-level covariates of overclaiming
Hypothesis 15: school-level rule violation and overclaiming	test motivated (positivity) bias hypothesis as a probable mechanism leading to overclaiming; gauge school-level covariates of overclaiming
Hypothesis 16, 17 & 18: socio-demographic correlates of overclaiming	expand knowledge on socio-demographic correlates of overclaiming (gender, socio-economic status, type of school, location size)
Hypothesis 19: latent class identification	explore individual differentiation of overclaiming and analyse covariates of overclaiming subtypes

Table 2. Hypotheses summary.

Chapter 6- DATABASE AND DATA PREPARATIONS

6.1 Basic information about PISA

6.1.1 What is PISA?

The Programme for International Student Assessment (PISA) is an international, large-scale, triennial assessment conducted both in the OECD and partner countries and regions. Last editions of PISA (2018 and the edition planned for 2021) covered data from around 90 entities, whereas the first study conducted in 2000 covered only 32 participating entities. The assessment is conducted on a group of 15-years-olds.

PISA is now widely used to examine national and cross-country disparities in learning outcomes, as well as factors related to learning and teaching practices across countries. The main assessment's goal is to monitor educational systems worldwide and provide methodologically-sound international comparisons. PISA items are designed to measure students' ability to use knowledge acquired in school to solve problems they might encounter in everyday life. An important purpose of PISA is also to provide evidence for temporal comparisons. Last but not least, PISA is also aimed at creating new scientific knowledge and informing educational policies worldwide with a special focus on assessment policies. Although some methodological aspects of this assessment have been criticized, PISA remains one of the most important scientific sources of information about education worldwide as in recent rounds of the PISA cycles around half a million students in more than 60 countries were tested. This makes PISA the largest educational research currently conducted.

Collected measures

Two main types of data collected in PISA are: a) cognitive tests, b) non-cognitive self-reports. The former entails three main domains: mathematics, reading (literacy in the main language of instruction in a given country) and science. Every edition is dedicated to one of the three domains which becomes main domain of the cycle on which special focus is given. Non-cognitive measures are also harmonised in content with the main theme of each edition. In example, in the 2012 cycle the background questionnaire was mainly composed of questions measuring mathematical abilities, attitudes and learning processes, whereas in the 2018 edition, when science was the point of focus, non-cognitive scales also predominantly measured school science-related topics. Apart from domain-related scales the PISA's non-cognitive part also comprises scales measuring school relations, attitudes and a set of socio-demographic factors, like parental education, socio-economic family status or immigrant background. Apart from students, the background information is also collected from school principals who fill in a web-based questionnaire (the so-called school background questionnaire). Teachers and parents can also fill in dedicated questionnaires but it is only available as an additional option (in the 2012 cycle Poland did not participate in neither of them).

Procedure and design

During the measurement sessions students participating in PISA are asked to first sit a cognitive test aimed at assessing their proficiency in reading, mathematics, and science. This part has a time limit of two hours. Then participants are to respond to questionnaire items in which they report on their attitudes, learning experience, motivation, family characteristics, etc. This part lasts for a maximum of one hour (normally 35-45 minutes). In each cycle countries can elect to participate in optional parts that vary from edition to edition. Optional assessments in the most recent cycles comprised e.g.

financial literacy measurement or collaborative problem-solving test. Countries can also chose to conduct measurement in additional, national add-ons assessments.

Both cognitive and non-cognitive measurements are organised in a rotated design (planned missing-data design), meaning that each respondent answers only to a subset of the whole set of items. Such subsets are called test forms. Students are randomly assigned to the test forms. This enables to balance content coverage, measurement precision and respondents' effort. More information on missing-data designs can be found in Pokropek (2011) and on PISA design specifically in Borgonovi and Biecek (2016).

PISA employs various kinds of item format in cognitive tests, apart from multiple-choice items also open-ended questions are used and this format typically comprises around one-third of the total number of items. Regarding non-cognitive scales predominantly Likert-type rating scales are used.

Mode of administration

PISA started as a paper-and-pencil assessment but since 2006 (in Poland- 2009) it gradually moved to computerised-based mode in order to become fully computerised from 2015[86]. However, also in previous cycles subsamples were selected to participate in computer-based assessments (Federowicz, 2013). The assessment is self-administered, grouped and proctored. Normally, it takes place within school's facilities.

Scaling and scoring

Probabilistic models from the IRT family are used to scale both cognitive and non-cognitive data in PISA.

Cognitive measures are scaled using the multidimensional (generalised) dichotomous Rasch model. Participants' scores are represented on the so-called PISA scale with a mean of 500 points and a standard deviation of 100 points. However, the scale is anchored in the PISA 2000 cycle and 500 points (the so-called PISA points) represents the average of the OECD countries in 2000 and 100 points is equal to one standard deviation in the same PISA 2000. This is done in order to guarantee longitudinal comparability of the PISA scores. It is noteworthy that the scale is calibrated for each general domain separately, so PISA 2012 mathematical scores are scaled to PISA 2003 mathematical scores as these are two PISA cycles in which math was the general domain (next time mathematics will be general domain in 2021). This methodological approach results in PISA scores being an interval-level measure, hence any ratios are not meaningful for this scale (more in Federowicz, 2013; OECD, 2014a; 2014b).

The main PISA goal is to provide inter-group comparisons and all its methodology is directed to achieve this objective. Therefore, individual-level scores are represented using the plausible values (PVs) methodology. In this method individual scores are drawn from a posterior distribution of probable scores which is calculated on the basis of test answers vector and a set of conditional variables. In PISA 2012 five PVs were generated per domain for each student. More on PVs use and characteristics can be found in the PISA 2012 technical report (OECD, 2014b) and in specialised publications (e.g. Von Davier, Gonzalez & Mislevy, 2009). PISA cognitive test scores were used in the form prepared by the OECD and no alterations were made to their scaling.

[86] In the 2015 cycle countries could opted for a paper-and-pencil assessment. Around one-fifth of the participating countries elected to do so, among them were only two from the European Union: Romania and Malta.

Non-cognitive measures in PISA 2012 can be divided into scales were items were just scored and no scale-level scoring was performed (e.g. gender) and scales were item scores were scaled in order to obtain indices. The latter group of item scores were scaled by the partial-credit model (PCM) which is a polytomous extension of the Rasch IRT model. The scaled scores are represented on the scale with an OECD average of zero and a standard deviation of one. These scales were rescaled and rescored in this work using the graded response model (GRM), a polytomous extension of the two-parameter logistic model (2-PLM). The new indices have a mean set at zero and a standard deviation of one within the Polish sample in order to ease interpretation of regression results.

Both students and school data are weighted in order to account for any inconsistencies between the sampling frame and factually assessed entities. Detailed comments about the PISA weighting system are well beyond the scope of this work but can be consulted in the PISA technical report for each cycle (e.g. OECD, 2014b). According to the recommendations formulated by Rutkowski, Gonzalez, Joncas and von Davier (2010) both school and student final weights are used in all analyses.

Organisation and data access policy

The measurement is organised and coordinated by the OECD, however it is paid from the national dues of the participating countries, typically these contributions come from the budget of a given Ministry of Education (MoE). The tests are conceived and prepared by a multinational consortium led by the OECD PISA team. Local data collection is organised by national teams and is supervised by the OECD.

Long before establishing the Open Access rules PISA followed similar requirements and since 2000 all data, measurement instruments and reports are freely downloadable from the OECD dedicated webpage. This excludes some of the cognitive items which serve as anchors linking to the previous PISA cycles and as such are used in every PISA cycle and cannot be revealed to the public.

6.1.2 Poland in PISA

Poland, member of the OECD since 1996, participated in every PISA edition, beginning from the first one conducted in 2000. Poland also eagerly participates in optional assessments, e.g. the country took part in all three editions of the financial literacy measurement. Optional ICT familiarity and educational career questionnaires are also frequently conducted among the Polish students.

The country initiated its PISA participation noting below OECD-average results in 2000, to soar to the group of top-performing countries, achieving scores well above the OECD-average in all three cognitive domains in the 2018 cycle (OECD, 2019). Despite the rising scores in cognitive test results Poland is notorious for reaching low scores in school climate and school belonging self-reports. Another interesting Polish characteristic is that it is among the countries that has one of the lowest gender gaps in cognitive PISA scores among the OECD countries. Socio-economic status is in Poland a slightly more important predictor of students' educational attainment than it is on average in the OECD countries. However, the inter-school equity is higher in Poland than in the OECD in general, with high- and low-performing students clustered in schools less than in many other countries[87].

Each country has an option to collect additional samples in order to perform inter-regional analyses (1500 students per region), however as so far Poland never elected to do so.

[87] Detailed information on the Polish participation in the PISA 2000-2015 cycles can be found in country reports prepared by the Polish MoE and the Polish PISA team led by Prof. Michał Federowicz. Other publications covering this topic is e.g. Jakubowski, Konarzewski, Muszyński, Smulczyk & Walicki (2017) report.

6.2 Database characteristics

6.2.1 Sample

Sampling process and response rate

PISA uses two-stage stratified samples of students enrolled in lower-secondary or upper-secondary institutions and aged between 15 years and 3 months and 16 years and 2 months in the year of the testing in order to represent the full population of this cohort in every participating country.

In the first stage of sampling a school-level sampling frame is constructed and validated. The sampling frame is stratified in order to cover important school characteristics (e.g. type, number of students enrolled, location size). Then, a random sample of a minimum of 150 schools is drawn to form an original school sample. For each school from this list two substitute (replacement) schools are drawn in order to guarantee replacements if a school from the original list refuses to participate. Response rate of 65% from the original list is required before schools from the replacement list can be recruited. In case of the PISA 2012 cycle in Poland only 12% of the schools measured were drawn from the replacement list (RR=88%; Federowicz, 2013). Totally, 184 schools took part in the Polish PISA 2012 core assessment[88]. The core school sample included 176 junior high-schools (*gimnazjum*) and seven other: two vocational schools and five high-schools.

The second stage of sampling comprises simple random sampling of students within each school. In order to participate in the standard PISA and in the optional financial literacy assessment (as did Poland in 2012) a minimum sample of 6300 students was required. No replacement samples were drawn as students cannot be replaced in PISA. Response rate of at least 80% was necessary for a given country data to be accepted by the OECD and reported in official publications. In 2012 Poland drew a total number of 6811 students for both PISA assessments (core and financial literacy) of which 5662 eventually completed PISA measurement tools yielding a response rate of 82%[89] (Federowicz, 2013). For the core PISA assessment 5545 students were drawn of which 4607 participated in measurements (RR=83%). Private school students were oversampled for the Polish assessment to a total number of 357 who actually took part in the core assessment.

As evidenced by the in-depth analyses conducted by the Polish PISA Team students' missingness was moderated by their ability level as low-achievers were more likely to skip the measurement for which they were drawn. It was estimated that this fact downbiased the PISA cognitive scores by a mere 3 PISA points (around 0,04 of a standard deviation; Federowicz, 2013). Due to the lack of data it is impossible to gauge how this missingness not-at-random influenced non-cognitive assessments' scores. However, low biases in case of cognitive assessments suggest that any bias in non-cognitive scales is most likely also extremely limited, in a range of a hundredth parts of standard deviation.

Rotation design

PISA background questionnaire was implemented in a rotation design (see section 6.2.2 below for more details), hence only a subsample completed math familiarity scale with overclaiming items embedded. Math familiarity scale was presented to 3071 students who were randomly drawn to Forms A and C of the questionnaire. Form B did not contain math familiarity scale on the basis of the PISA missing-by-design structure.

[88] One additional school participated in the financial literacy test only.

[89] Precise response rates (RR) often marginally differ between the documents as slightly different RR counting procedures are used by different reporting entities.

However, the number of participants possible to include in analyses is slightly lower due to item non-response. The table below summarises missing data by its cause:

Item	Number of missing		% of missing	
	Missing-by-design	Missing	Missing-by-design	Missing by non-response
st62q01	1536	45	33,34%	1,47%
st62q02	1536	40	33,34%	1,30%
st62q03	1536	41	33,34%	1,34%
st62q04	1536	42	33,34%	1,37%
st62q06	1536	42	33,34%	1,37%
st62q07	1536	31	33,34%	1,01%
st62q08	1536	47	33,34%	1,53%
st62q09	1536	33	33,34%	1,07%
st62q10	1536	43	33,34%	1,40%
st62q11	1536	50	33,34%	1,63%
st62q12	1536	32	33,34%	1,04%
st62q13	1536	39	33,34%	1,27%
st62q15	1536	37	33,34%	1,20%
st62q16	1536	34	33,34%	1,11%
st62q17	1536	33	33,34%	1,07%
st62q19	1536	27	33,34%	0,88%

Table 3. Missing data in each math familiarity item by type of missing. Missing-by-design % was calculated from the whole eligible sample (4607), while missing by item non-response was calculated from participants that could respond to an item (3071).

Moreover, number of missing values yielded in a given math familiarity scale factor was calculated. The results are displayed in Table 3 below:

Number of missing values	Number of students	Percent
0	2881	93,81%
1	123	4,01%
2	33	1,07%
3	1	0,03%
4	3	0,10%
5	2	0,07%
6	1	0,03%
7	2	0,07%
10	2	0,07%
11	1	0,03%
15	1	0,03%
Total	3071	100%

Table 4. Frequency of number of missing values in the math familiarity scale vector.

The analysis showed that most of the subsample did not yield any missing values. Most of the participants with data missing did not respond only to one or two items. However, there was also a participant who answered only one item out of the 16 in the math familiarity scale.

Missing data

Because different participants noted a minor proportion of missing data on different scales the precise sample size differs in some analyses. In general either 3071 or 2881 unique response vectors was used in each analysis. The lower number of respondents was used when missing data was not allowed by software requirements (e.g. see subchapter 7.5).

It was decided not to impute the missing data basing on: a) low percent of missing data, close to the 1% threshold termed as "excellent" (Borgonovi & Pokropek, 2019), b) low certainty of what variables should be included in the multiple imputation equations due to exploratory character of many analyses and insufficient coherence of the previous evidence on math familiarity and OCT correlates, c) probable low influence of imputation on models results, due to c') missing-by-design data treated as missing completely as random (MCAR) and c'') low fraction of data missing due to item non-response, d) a vast majority of the missing data is classified as MCAR, hence any estimates using this data are unbiased due to data missingness, e) including imputed data, especially from the missing-by-design part (Form B) would probably not change estimates' results (e.g. coefficients' values) but would only downsize standard errors due to increased statistical power. Such increase would reduce probabilities of inferential errors, however, in this work most of the results will be treated conservatively anyway due to initial character of many analyses. What is more, the remaining sample size is large anyway and secures sufficient statistical power for the analyses presented here. Similar decisions were made by e.g. Reise et al. (2016) and Vonkova et al. (2018).

Basic sample characteristics

Among the students that were presented the math familiarity questionnaire 99.4% was a junior high school student, almost all of them in the 3rd (final) grade of this educational level. Female students consisted slightly more than half of participants (51.7%). Foreign-born students consisted only 0.4% of the subsample. Other sample characteristics will be given when relevant to the substantial analysis as presented in the subsequent chapter.

6.2.2 Materials

Rotation design (missing-by-design) of the PISA 2012 background questionnaire

In the 2012 PISA cycle only the core socioeconomic and demographic background questions were administered to all students. The content of this part is briefly presented in the table below:

Item	Description
st01	Grade attending
st02	Country
st03	Age of student
st04	Sex of student
st05	Attending preschool
st06	Attending kindergarten
st07	Grade repeating
st08, st09, st115	Truancy
st11	Family structure
st12-st19	Parents' educational and occupational status
st20, st21, st25	Immigrant background
st26-st28	Household possessions

Table 5. Schematic presentation of the content of the PISA 2012 student questionnaire common (core) part. All students sitting PISA were presented these questions.

The rest of the background (student) questionnaire followed a rotation design such that only two thirds of the overall sample of students received, at random, any one question (OECD, 2014b). The rotation design was adopted to increase the overall amount of topics that could be explored without increasing the response burden for individual students. The design is schematically represented in the table below:

Form A		Form B		Form C	
Item	**Description**	**Item**	**Description**	**Item**	**Description**
st01-st28	common part	st01-st28	common part	st01-st28	common part
st29, st35, st37, st43, st44, st46, st48, st49, etc.	part A	st42, st77, st79, st80-st91	part C	st53, st55, st57, st61, st62, st69, st70-st76	part B
st53, st55, st57, st61, st62, st69, st70-st76	part B	st29, st35, st37, st43, st44, st46, st48, st49, etc.	part A	st42, st77, st79, st80-st91	part C

Table 6. Schematic representation of the rotated design of the PISA 2012 student background questionnaire.

Full information on items and scales forming given parts can be found in the PISA 2012 technical report (OECD, 2014b, p. 61).

Math familiarity scale

Key analyses that follows in the subsequent chapter concentrate on math familiarity scale (st62[90]) and the embedded overclaiming items. As evidenced by Table 5 above this scale was included in part B of the student background questionnaire and hence was not presented to students that were randomly drawn to complete Form B of the background questionnaire which contained only parts A and C. This reduces the overall sample size for all analyses using st62 scale by approximately one-third.

The scale comprises 16 items altogether, including 13 reals and three foils. Participants responded on a five-categorical, Likert-type response scale, labelled with both numbers and descriptions, from "Never heard of it" (1) to "Know it well, understand the concept" (5). Respondents were asked to tick only one answer (box) in each row. Precise wording of the scale's stem, items and response categories are presented in the table below[91]:

[90] Each background question is coded by an international number. This helps to orientate between the items that are presented in different questionnaire forms and in different language versions. Letters "st" stand for an item from the student background questionnaire, whereas "sch" code items presented in the school principals inventory. If items in a scale are not in mathematical order, e.g. some item numbers are missing, it means that these items were deleted after the field trial pilot study. If items are coded with a high number, e.g. above 110, it usually means that the items were profoundly revised after the trial (OECD, 2014b). Such identification numbers will be always used in this work in order to avoid confusion regarding which scale is being analysed.

[91] Items in the Polish version are taken from the Polish version of the PISA 2012 background questionnaire. In the paper-and-pencil questionnaire the scale's items were presented in a fixed order that is imaged in Table 3.

Thinking about mathematical concepts: how familiar are you with the following terms?							
Item	**English**	**Polish**	**Never heard of it**	**Heard of it once or twice**	**Heard of it a few times**	**Heard of it often**	**Know it well, understand the concept**
st62q01	exponential function	funkcja wykładnicza	1	2	3	4	5
st62q02	divisor	dzielnik	1	2	3	4	5
st62q03	quadratic function	funkcja kwadratowa	1	2	3	4	5
st62q04	*proper number*	*liczba właściwa*	1	2	3	4	5
st62q06	linear equation	równanie liniowe	1	2	3	4	5
st62q07	vectors	wektory	1	2	3	4	5
st62q08	complex number	liczby zespolone	1	2	3	4	5
st62q09	rational number	liczba wymierna	1	2	3	4	5
st62q10	radicals	pierwiastek	1	2	3	4	5
st62q11	*subjunctive scaling*	*skalowanie łączące*	1	2	3	4	5
st62q12	polygon	wielokąt	1	2	3	4	5
st62q13	*declarative fraction*	*ułamek oznajmujący*	1	2	3	4	5
st62q15	congruent figure	figura przystająca	1	2	3	4	5
st62q16	cosine	cosinus	1	2	3	4	5
st62q17	arithmetic mean	średnia arytmetyczna	1	2	3	4	5
st62q19	probability	prawdopodobieństwo	1	2	3	4	5

Table 7. Math familiarity scale (st62) in Polish and English version with international item numbers in the first column. Foils are in italics.

According to the PISA 2012 technical report foils were created by combining a real grammar term (e.g. "proper" or "declarative") with a real mathematical term as to form a foil item that in its entirety does not mean anything. This method of foil creation would be classified as yielding foils of high risk of confusion with existing mathematical concepts (cf. Franzen & Mader, 2019; Hargittai, 2005; Paulhus et al., 2003).

Reals employed in the scale mainly stem from algebra and geometry. However, efforts to form algebra familiarity and geometry familiarity indices were not finalised[92] (OECD, 2014b).

Other scales and materials used

The additional scales and indices will be commented in the relevant sections of Chapter 7.

[92] It is also advisable to consult discrepancies in item numeration between the technical report and field questionnaire version on page 57 of the report (OECD, 2014b).

6.2.3 Data preparation

Data used

In the analyses that follow data from three sources were used: a) PISA cognitive items, namely mathematical test, b) indices from the PISA background questionnaire, with a special focus on the math familiarity scale, c) data from the school questionnaire, where answers were submitted by school principals. School-level data was merged with the student-level dataset for that purpose.

Math familiarity scaling and scoring

Before scaling and scoring basic psychometric properties of the scale were examined. The analysis presented here intended only to verify scale's internal consistency, calculated by Cronbach's alpha and a handful of basic psychometric measures. The analysis did not reveal any serious problems with scale coding, internal consistency, outlying items or suggestions of dimensionality problems. The topic of internal structure of math familiarity scale will be revisited in more detail in the subsequent chapter. The analyses conducted here are summed up in the table below:

Item	N	Sign	Item-test correlation	Item-rest correlation	Average interitem covariance	Alpha
st62q01	3026	+	0,57	0,49	0,39	0,851
st62q02	3031	+	0,62	0,54	0,39	0,848
st62q03	3030	+	0,63	0,54	0,37	0,848
st62q04	3029	+	0,64	0,56	0,38	0,847
st62q06	3029	+	0,63	0,55	0,38	0,848
st62q07	3040	+	0,58	0,48	0,38	0,851
st62q08	3024	+	0,59	0,51	0,39	0,850
st62q09	3038	+	0,63	0,57	0,39	0,848
st62q10	3028	+	0,53	0,48	0,41	0,853
st62q11	3021	+	0,45	0,36	0,40	0,857
st62q12	3039	+	0,55	0,49	0,41	0,852
st62q13	3032	+	0,49	0,39	0,39	0,857
st62q15	3034	+	0,61	0,53	0,38	0,849
st62q16	3037	+	0,49	0,39	0,40	0,856
st62q17	3038	+	0,58	0,52	0,40	0,850
st62q19	3044	+	0,57	0,50	0,39	0,850
Scale	-	-	-	-	0,39	0,859

Table 8. Internal consistency of math familiarity scale.

Another very basic analysis conducted here regarded raw data descriptive statistics, namely frequencies of response categories embraced:

Item	English	Never heard of it	Heard of it once or twice	Heard of it a few times	Heard of it often	Know it well, understand the concept
st62q01	exponential function	10,60%	18,70%	29,70%	26,50%	14,40%
st62q02	divisor	3,40%	7,70%	12,60%	21,60%	54,80%
st62q03	quadratic function	15,90%	18,40%	20,50%	19,90%	25,40%
st62q04	*proper number*	9,50%	16,00%	24,00%	27,50%	22,90%
st62q06	linear equation	20,00%	20,20%	23,20%	20,90%	15,80%
st62q07	vectors	16,30%	18,40%	21,30%	22,50%	21,60%
st62q08	complex number	40,90%	25,30%	18,10%	11,00%	4,80%
st62q09	rational number	2,50%	7,00%	15,80%	30,50%	44,30%
st62q10	radicals	1,40%	2,60%	5,40%	17,50%	73,10%
st62q11	*subjunctive scaling*	54,60%	19,10%	13,50%	7,70%	5,10%
st62q12	polygon	1,40%	2,90%	6,70%	16,80%	72,10%
st62q13	*declarative fraction*	30,10%	20,60%	19,80%	15,50%	14,10%
st62q15	congruent figure	6,60%	8,70%	15,60%	22,90%	46,20%
st62q16	cosine	39,50%	24,50%	17,70%	11,50%	6,80%
st62q17	arithmetic mean	1,80%	4,10%	9,90%	18,50%	65,70%
st62q19	probability	3,20%	7,60%	13,30%	23,10%	52,70%
Mean reals		12,58%	12,78%	16,14%	20,25%	38,28%
Mean foils		31,40%	18,57%	19,10%	16,90%	14,03%
Mean total		16,11%	13,86%	16,69%	19,62%	33,74%

Table 9. Math familiarity scale- frequencies of response categories.

The above table reveals that scale's items differed greatly in their familiarity claimed by students. The most recognised items were "polygon" and "radicals", closely followed by "arithmetic mean". "Exponential function" and "linear equation" can be counted among items that were moderately claimed, whereas "cosine" or "complex number" were hardly familiar at all. Among the foils "proper number" was definitely the most alluring as more than 20% of students (over) claimed being very familiar with this item.

The averaged differences between reals and foils are imaged in the figure below:

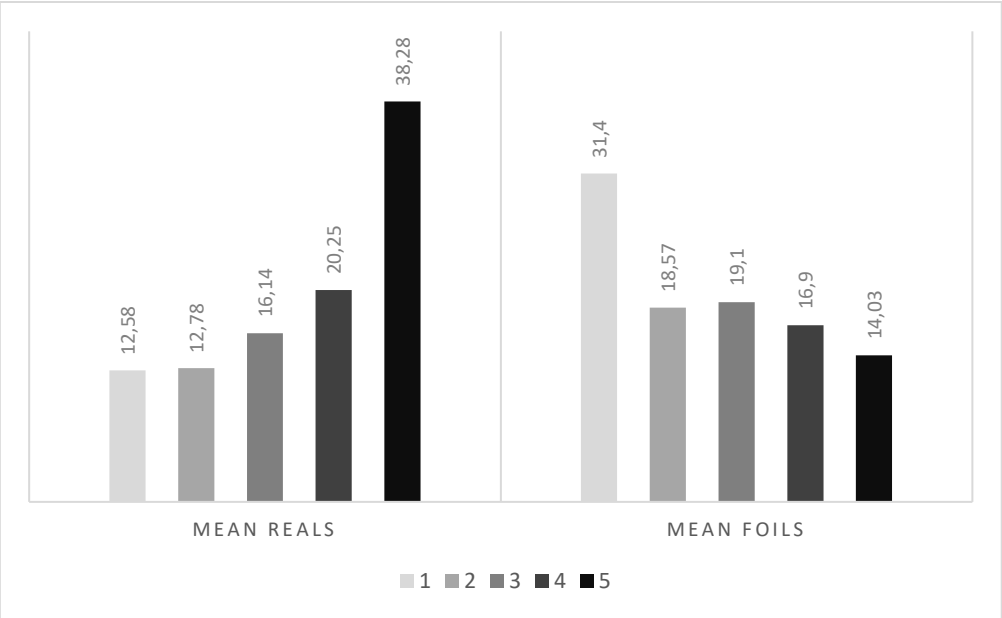

Figure 5. Difference between reals and foils claimed familiarity

The above figure shows that on average responses to reals where tilted towards claiming familiarity with mathematical concepts, whereas on average foils where claimed much less than reals. It is noteworthy that main difference between two kinds of items involves extreme response categories (1 and 5) as mean frequencies of using mid responses (2, 3, 4) do not differ much between reals and foils.

According to the analysis performed by Vonkova and co-workers (2018) Poland was classified as a country where students claimed lower than average familiarity with math concepts and overclaimed above the average of the countries participating in the PISA 2012 cycle.

Math familiarity scale with the embedded overclaiming items was rescaled and rescored in order to yield indices more comparable with the rest of the OCT literature (e.g. Paulhus et al., 2003). Signal detection theory (SDT) indices were calculated on the basis of the math familiarity scores according to the procedure described by Paulhus and Petrusic (2010) and Vonkova and colleagues (2018).

Signal Detection Theory (SDT) scoring

Signal detection theory was born during the troubled years of the Second World War during which it was developed as a theoretical underpinning for radar usage. The theory first used by physicists and mathematicians very soon found its application in social sciences, chiefly psychology. In this field SDT was employed to describe various cognitive operations, mainly perceptual, mnemonic and decision-making processes.

The theory resides on a matrix of two possible states of signal (target, stimulus) that can be present and absent and two decisions that can be made regarding this signal- respond that it is present (response “yes”) or respond that it is absent (decision “no”). The below table sums up possible combinations and consequences of decisions in a given signal state:

Decision	Signal present	Signal absent
"YES"	hit (H)	false alarm (FA)
"NO"	miss (M)	correct rejection (CR)

Table 10. Matrix of SDT decisions and four possible results. Results' abbreviations in brackets ().

Among the four basic results of SDT are hit (H), when stimulus was indeed present and it was correctly identified by responder, false alarm (FA) when stimulus was absent but respondent made an incorrect decision of answering that it was present. In the realities of OCT an answer would be classified as a hit when high response categories (e.g. 4 or 5) would be used to an existing mathematical concept (e.g. "polygon"), whereas a false alarm would occur when such response categories would be used when responding to a non-existing concept (e.g. "subjunctive scaling").

On the other hand, an error of omission or miss (M) is committed when signal is present but respondent fails to distinguish it from the noise. In the world of survey responses such situation can occur when respondent would claim not to know an existing mathematical concept (e.g. "cosine"). Lastly, the correct rejection's (CR) name speaks for itself- it takes place whenever respondent correctly identifies lack of signal as such (noise). In the OCT framework it occurs whenever respondent uses response categories coding lack of knowledge when assessing a non-existing concept (e.g. "declarative fraction"). Paulhus and co-authors advocated use of SDT in scoring OCT results as SDT enables to jointly model maximum of the relevant information from the task, including modelling decisions' sensitivity and specificity (2003). This is an advancement in comparison to other scoring rules, e.g. using simple sum of responses to foils as an index of overclaiming (see e.g. Hulur et al., 2011; Steger et al., 2020). Moreover, SDT measures are based on information from all the items, hence they are more reliable than indices based on just reals or foils (such as sum of scores on foils) as they are based on more information stemming from a larger number of items (Paulhus et al., 2003).

Therefore, in the frames of SDT a rich variety of indices can be calculated on the basis of decision matrix presented in Table 7. Potential number of such indices is vast but here they will be limited to presenting two indices used in this work: *d'* and *c*. These indices were selected due to recommendations of Paulhus and Petrusic (2010) who compared validity of several indices in the OCT context. Moreover, the above-mentioned indices are also often used in the OCT research, hence their implementation enables greater results' comparability across studies.

The first of the two indices, *d'*, is a sensitivity index coding how well a given respondent distinguishes between reals and foils (signal and noise). The higher its value the better the discrimination. It is a dimensionless statistic based on assumption that signal and noise variances are equal and that both are normally distributed. There are various ways to calculate it, the easiest one is defined in the equation below:

$$\boldsymbol{d' = Z(PH) - Z(PFA)} \text{ (Equation 1)}$$

where „Z" is the inverse of the cumulative distribution function of the Gaussian distribution (Φ^{-1}), "PH" is hit rate (proportion of hits to sum of hits and omissions) and "PFA" is false alarm rate (proportion of false alarms to sum of false alarms and correct rejections). In case of items with polytomous scores,

e.g. rating scales, the situation is a bit more complicated as the simple dichotomous logic of SDT cannot be readily applied to Likert-type ratings. To extend SDT to such scores PH and PFA are thus calculated on a given threshold (cut-off point) or averaged across all thresholds. For example, in a rating scale like the one used in the PISA 2012 math familiarity scale there are five response categories, hence there are four cut-off points (thresholds) (1-2, 2-3, 3-4, 4-5) on which PH or PFA can be calculated (see details in Vonkova et al., 2018).

The second of the indices, *c*, also named criterion location or decision criterion (often shortened to criterion), is a measure of predilection toward responding rather affirmatively to the items or rather negatively. The index can be calculated as follows:

$$\boldsymbol{c} = -\frac{\boldsymbol{Z(PH)+Z(PFA)}}{\boldsymbol{2}} \text{ (Equation 2)}$$

The measure shows what is considered by a given respondent as signal and what as noise. In the social sciences, e.g. in the OCT framework, it is used as a measure of bias. Negative values of criterion means liberal criterion, in the framework of OCT propensity to claim familiarity with foils and reals alike (answering "yes"). Values close to zero code neutral criterion (no bias)[93], whereas high values mean conservative criterion, propensity to answer "no". However, the minus sing in Equation 2 is often omitted so that high scores indicate bias towards affirmative responding and low scores bias towards negative responding (Stanislaw & Todorov, 1999). This version of this statistic, the so-called c-reversed will be used in this work.

Unlike many others indices *d'* and *c* are dimensionless (independent of measuring scale) and are calculated independently from each other, though they are predominantly correlated across individuals (Paulhus et al., 2003).

More information on SDT indices can be found e.g. in the articles written by Abdi (2007) and Stanislaw and Todorov (1999) and in comprehensive, book-length positions (e.g. Macmillan & Creelman, 1991/2005; McNicol, 1972/2005).

Oftentimes, the so-called "common sense" indices are employed which mainly reside on simplifying the equations for SDT indices by not calculating z-scores. A pair of such indices was introduced by Paulhus et al. (2003) and Vonkova et al. (2018): *d'* is replaced by "index of accuracy" (IA) calculated as:

$$\textbf{IA = PH – PFA} \text{ (Equation 3),}$$

whereas criterion is supplanted by "index of exaggeration" (IE) calculated as (minus sign is also often omitted as in case of *c*):

$$\textbf{IE = -(PH + PFA)/2} \text{ (Equation 4).}$$

[93] *c* = 0 indicates an "ideal observer", a completely unbiased respondent (Abdi, 2007). The *c* is also often interpreted as a distance between actual strategy (threshold, criterion) and an ideal strategy.

IA and IE averaged across all the cut-off points are called averaged index of accuracy/exaggeration (IA^a/IE^a) but for the sake of simplicity IA and IE will be used in the sense of averaged indices in this work. IA values are in the interval [-1, 1] with the maximum of 1 indicates that all reals were responded to by category "5" (maximum familiarity) and all foils by "1" (no familiarity), whereas the minimum of -1 indicates that all reals got "1" and all foils "5". If IA=0 than reals and foils were given the same ratings of familiarity. IE values are in the interval [0, 1], value 1 indicates that all items are rated on a maximum category ("5"), value 0- that all items are rated on a minimum category ("1") and value of IE of 0.5 means that all items were rated using the midpoint of the scale ("3").

In order to maintain accord between the three measures of OCT bias used in the subsequent analyses *c* and IE were reversed. In this line the interpretation of *c*, IE and an index of overclaiming calculated from foils directly have the same interpretation: the higher the score, the higher the propensity to answer "yes", namely, use "4" and "5" response categories in the math familiarity scale.

IRT scaling

Math familiarity scale scores were scaled using IRT partial credit model (PCM) by the OECD PISA team, however on the basis of few premises it was decided to rescale the math familiarity scores using IRT graded response model (GRM) due to: a) PCM index used metric where OECD mean equalled zero and OECD standard deviation was equal to one (OECD, 2014b, p. 312), which was considered useless and even unconstructive for a one-country analysis, b) almost certain better model fit of a GRM model over a PCM model (Maydeu-Olivares, 2005, 2015)[94], c) want to generate math familiarity scores using PVs in order to c') have matching for math ability scores represented by five PVs and also c'') to obtain more precise estimates in comparison to point estimates used originally in the PISA data[95] (Rutkowski et al., 2010) and d) want to generate separate scores for reals and for foils as the OECD-generated indices did not include an index representing foils information only.

To calculate GRM Stata module (package) -uirt- (Kondratek, 2016/2020) was used[96]. The basic IRT item characteristics of the math familiarity scale are presented in the table below:

Item	Coefficient	Robust S.E.	Item	Coefficient	Robust S.E.
st62q01			st62q06		
A	1,08	0,05	a	1,01	0,04
b1	-2,37	0,10	b1	-1,66	0,08
b2	-1,08	0,06	b2	-0,58	0,05
b3	0,30	0,04	b3	0,52	0,05
b4	1,90	0,08	b4	1,87	0,09
st62q02			st62q07		
A	2,22	0,08	a	1,12	0,05
b1	-2,29	0,08	b1	-1,82	0,08
b2	-1,55	0,05	b2	-0,80	0,05
b3	-0,94	0,04	b3	0,14	0,04
b4	-0,23	0,03	b4	1,30	0,06

[94] Performed comparison of AIC and BIC values pointed to the indeed superior fit of the GRM solution.
[95] Weighted Likelihood Estimation was employed to obtain individual scores (OECD, 2014b, p. 312).
[96] The estimates were compared with Stata -irt- module and both estimates were essentially the same.

st62q03			st62q08		
A	1,12	0,05	a	0,78	0,04
b1	-1,86	0,08	b1	-0,56	0,06
b2	-0,81	0,05	b2	0,92	0,07
b3	0,12	0,04	b3	2,35	0,13
b4	1,10	0,06	b4	4,25	0,23
st62q04			st62q09		
A	1,23	0,05	a	2,18	0,08
b1	-2,25	0,09	b1	-2,53	0,09
b2	-1,16	0,05	b2	-1,67	0,05
b3	-0,12	0,04	b3	-0,88	0,03
b4	1,17	0,06	b4	0,11	0,03
st62q10			st62q15		
A	3,18	0,14	a	1,87	0,07
b1	-2,44	0,08	b1	-2,04	0,07
b2	-1,91	0,06	b2	-1,42	0,05
b3	-1,43	0,04	b3	-0,75	0,03
b4	-0,71	0,03	b4	0,04	0,03
st62q11			st62q16		
A	0,44	0,04	a	0,71	0,04
b1	0,43	0,09	b1	-0,71	0,07
b2	2,37	0,22	b2	0,86	0,07
b3	4,41	0,40	b3	2,27	0,13
b4	6,78	0,61	b4	4,05	0,23
st62q12			st62q17		
A	3,25	0,15	a	2,88	0,12
b1	-2,42	0,08	b1	-2,41	0,08
b2	-1,89	0,06	b2	-1,79	0,05
b3	-1,33	0,04	b3	-1,15	0,04
b4	-0,68	0,03	b4	-0,51	0,03
st62q13			st62q19		
A	0,57	0,04	a	1,82	0,07
b1	-1,62	0,12	b1	-2,60	0,10
b2	-0,03	0,07	b2	-1,72	0,06
b3	1,54	0,12	b3	-0,99	0,04
b4	3,31	0,23	b4	-0,15	0,03

Table 11. Math familiarity scale item characteristics under GRM model. Note: S.E.- standard error.

Estimated item characteristics confirm overall acceptable fit of the model as a) all discrimination parameters are above 0, though not all of them are above 1, suggesting poor item qualities or certain item-fit problems and b) all slopes are ordered in the predicted direction. It is also evident that items differ widely regarding their difficulty. Visual inspection of item characteristic curves (category characteristic curves, CCC) indeed shows that expected model does not fit well to the observed data in case of some items. The main cause of such misfit was salient already in the frequency table- some items were answered using almost exclusively two response categories: 1 or 5. The topic will be revisited later on. The items' CCCs are available in Online Appendix A.

PVs scoring

In addition to the previously calculated scores of the math familiarity scale (PH, PFA, IA, IE, *d'*, *c*) IRT scoring was also performed. Separate scores were calculated for reals and foils using uirt Stata module with the use of the GRM. Five PVs were generated for each of the item types. The PVs were conditioned on participants' gender, socio-economic status (PISA escs index was used here, see OECD, 2014b), math abilities and school clustering in order to enhance precision of point estimates performed with the use of PVs[97] (Rutkowski et al., 2010; Wu, 2005).

Whenever PVs were used in any analysis the magnitude of coefficients was averaged across imputations for PVs and standard errors were adjusted to the sampling design and to imputation variability by using proper weighting and Rubin's rules (Jerrim, Lopez-Agudo, Marcenaro-Gutierrez & Shure, 2017; OECD, 2014b; Rubin, 1996).

Other scales

All other self-report data where computing statistical indices was applicable was rescaled using uirt Stata module. The questionnaires were scaled by the GRM and scored using the Expected a Posteriori method to estimate individual scores and their standard errors. Such estimates should typically yield correct point estimates, but their standard errors would be underestimated in comparison to generating PVs (Rutkowski et al., 2010; Wu, 2005).

The scaling procedure was preceded by reliability, inter-item polychoric correlations and dimensionality analysis in order to browse the scales for any trouble-making items and problems with dimensionality. The detailed results of these analyses are presented in Online Appendix B. Whenever directly relevant to the matters at hand this issue will be also commented in the next chapter.

Software

Various statistical packages where used throughout the analyses performed in this work. Information on precise package employed is presented in the relevant sections of the subsequent chapter. In general, R (version 4.0.0), Stata (version 14.2 SE) and MPlus (version 8.2) were used.

Results presentation

Whenever statistical significance is presented in a table the following coding system is adopted: *p*< 0.001- no sign[98], *p*< 0.01 **, *p*< 0.05 *, ns- non-significant.

The subsequent chapter presents and comments the verifications of the hypotheses formulated in the chapter 5. The presentation of results is organised as follows: at first short methods section is presented, commenting on specific sample, material or statistical features of a given analysis, then results are displayed and finally short discussion concludes each hypothesis part.

[97] Unconditioned PVs tend to yield estimates biased towards zero, namely regression coefficients smaller than they should be (Wu, 2005).

[98] Conventionally, three stars (***) are produced when *p*< 0.001 but in this work the stars are omitted in this particular case as due to the large sample size most of the coefficients are either non-significant or significant at the *p* level of 0.001.

Chapter 7- RESULTS OF THE HYPOTHESES TESTING

7.1 Overclaiming scores as a suppressor of spurious variance- Hypothesis 1

7.1.1 Method

In this analysis the suppression hypothesis will be tested. It assumes that OCT bias score will act as a classical suppressor for the relation between math self-report and math ability test (criterion). To this end a multilevel regression was performed with math ability test results as dependent variables and various measures of self-report as independent variable. Moreover, measures of overclaiming bias were introduced into the equation to observe whether any suppression relation takes place (Conger, 1974).

To account for the PISA complex data structure a multilevel regression with correction for multiply imputed measures of student skills was performed in order to estimate unbiased measurement errors. To this end user-written Stata modules "pv" (MacDonald, 2014) and "repest" (Avvisati & Keslair, 2017) with Stata command -mixed- and -mi estimate- were used. Sampling was applied to both levels of analysis: PISA final student weights were used on the student-level while scale-adjusted total student weights (sum of final student weights in each school)[99] were used on the school-level (OECD, 2014b).

The PISA math ability test scores represented by five PVs were used as dependent variable (criterion). In the role of independent variable different measures of math familiarity (OCT accuracy) were used: PVs calculated on the basis of reals, *d'* and index of accuracy. These measures were used as self-reported math ability. A positive and relatively high relation between this measure and math ability was expected. As suppressor variable different measures of overclaiming bias were used: PVs calculated on the basis of foils, *c* and index of exaggeration (both reversed, see 6.3.2). Each pair of the indices was used in a separate regression.

In the first step zero-order correlations between key variables were calculated. Afterwards, a series of regression equations was performed, when in the first step an independent variable was introduced, in the second step a suppressor was introduced and in the third and final step both variables were introduced together in order to examine whether suppression effect takes place. Interaction in the fourth step was added in order to check for any moderation effects. Change in the magnitude of the regression coefficient between dependent and independent variable would be the key indicator of whether suppression truly takes place.

[99] The discussion about weighting designs suitable for multilevel analyses of the PISA data is ongoing (e.g. Asparouhov, 2006; Avvisati, 2020). In most of the cases using any of the correct weighting designs yields very similar results (Jerrim et al., 2017). In case of the suppression analysis presented in this work three systems were compared and they all yielded essentially the same results.

7.1.2 Results

The table below presents zero-order correlations for the key variables used in the regression equation:

Variable	**math ability**	**math familiarity (reals)**	**bias (foils)**	***d'***	**c reversed**	**i_accuracy**	**i_exaggeration _reversed**
math ability	1						
math familiarity (reals)	0,57	1					
bias (foils)	-0,11	0,13	1				
d prime	0,34	0,27	-0,60	1			
c reversed	0,16	0,54	0,65	-0,54	1		
i_accuracy	0,41	0,39	-0,61	0,90	-0,37	1	
i_exaggeration _reversed	0,21	0,61	0,65	-0,39	0,95	-0,32	1

Table 12. Zero-order correlations for math ability, OCT accuracy and OCT bias measures. Note: All ps < 0.001.

The analysis of Table 11 shows that all OCT accuracy indices correlate positively and quite highly with math ability. The largest pairwise correlation coefficient was yielded by IRT-scaled and PV-scored index of math familiarity, followed by SDT-derived measures of *d'* and index of accuracy. The correlation between math ability and self-reported familiarity with math concepts (proxy of math ability) of r=0.57 is quite high, especially in the light of notoriously low correlations between self-reports and their objective criteria (see subchapter 4.2).

Among the measures of bias IRT-scaled and PV-scored index of bias correlated negatively with math ability, whereas SDT indices, *c* and index of exaggeration correlated positively, indicating that the more liberal the respondents were (more predilection toward answering "yes" in OCT) the higher their math ability. These results will be revisited in the section 7.2.2 below.

Let's now move to the regression equations. For the ease of interpretation and comparison between the tables the SDT (*d'*, *c*) and common sense indices (IA, IE) were standardised for a mean of zero and standard deviation of one to match the scaling of IRT indices. The R^2 were generated using -mlt- Stata package (Moehring & Schmidt, 2013); Raudenbusch-Bryk type of this measure was used (2002).

Parameter	Model 1		Model 2		Model 3		Model 4	
	B	p>z	B	p>z	B	p>z	B	p>z
math familiarity (reals)	46,34	***	-	-	49,63	***	49,96	***
bias (foils)	-	-	-4,91	*	-14,92	***	-14,64	***
interaction: reals*foils	-	-	-	-	-	-	-2,80	ns
R2 Level1		0,28		0,01		0,31		0,31
R2 Level2		0,49		0,03		0,57		0,57

*Table 13. Suppression analysis for PV-scored math familiarity scale. Note: ns p> 0.05, *p<0.05, **p<0.01, ** * p<0.001. B- regression weights, PISA scale where 1SD=100.*

Parameter	Model 1		Model 2		Model 3		Model 4	
	B	p>z	B	p>z	B	p>z	B	p>z
d’	23,73	***	-	-	46,42	***	54,77	***
c reversed	-	-	13,01	***	38,65	***	37,96	***
interaction: d’*c reversed	-	-	-	-	-	-	8,41	***
R2 Level1		0,07		0,02		0,21		0,23
R2 Level2		0,25		0,02		0,45		0,48

*Table 14. Suppression analysis for SDT-scored math familiarity score. Note: ns p> 0.05, *p<0.05, **p<0.01, ** * p<0.001. B- regression weights, PISA scale where 1SD=100.*

Parameter	Model 1		Model 2		Model 3		Model 4	
	B	p>z	B	p>z	B	p>z	B	p>z
i_accuracy	28,66	***	-	-	42,45	***	42,73	***
i_exaggeration (reversed)	-	-	16,97	***	33,62	***	34,33	***
interaction: i_accuracy*i_exaggeration (reversed)	-	-	-	-	-	-	1,60	ns
R2 Level1		0,11		0,04		0,24		0,24
R2 Level2		0,29		0,04		0,45		0,45

*Table 15. Suppression analysis for common sense-scored math familiarity scale. Note: ns p> 0.05, *p<0.05, **p<0.01, ** * p<0.001. B- regression weights, PISA scale where 1SD=100.*

The suppression analysis results for all three systems of coding OCT are presented in the table below. Parameter change (B or R^2) is defined as difference between the value from Model 3 and from Model 1 for predictor change and Model 2 for suppressor change. Interaction term is taken from Model 4.

Instead of *p* values for the significance of parameter change the percentage change was given as all the parameters differences were significant at least at the 0.05 level[100].

Index type		predictor B change	suppressor B change	interaction term	ΔR^2 level1	ΔR^2 level2
IRT indices	parameter	3,29	10,91	-2,80	0,03	0,08
	%	7,10	203,87	ns	10,71	16,33
SDT indices	parameter	22,69	25,64	8,41	0,14	0,20
	%	95,62	197,08	***	200,00	80,00
common sense indices	parameter	13,79	16,65	1,60	0,13	0,16
	%	48,12	98,11	ns	118,18	55,17

Table 16. Change of B and R^2 parameters between Model 3 and Model 1. B- regression weight, PISA scale where 1SD=100.

7.1.3 Discussion

Conducted analyses pointed that indeed the suppression relation takes place in case of all three OCT scoring systems. It is to say, including a measure of overclaiming leads to enhancement of the predictive validity of self-report scale measuring academic skill. This enhancement is displayed by the increased regression weights and the rising value of R^2.

Among the three scoring systems SDT indices and common sense indices led to a much stronger suppression effect than IRT indices. This is predictable from a much higher correlation between *d'* and *c* (or IA and IE) than between IRT scores for reals and for foils, as correlation between predictor and suppressor is a necessary precondition for suppression effect to take place (Conger, 1974; Lancaster, 1999; Paulhus, Robins, Trzesniewski & Tracy, 2004). Moreover, the recommendations formulated by Paulhus and colleagues (2003), that whenever SDT indices are used in a regression equation both indices (accuracy and bias/discrimination and criterion) has to be included lest the results be severely distorted, where corroborated in the above analysis. In case of IRT-derived indices the consequences of excluding a suppressor from the analysis are not as pronounced but most certainly including it boosts the predictive validity of both independent variables and enhances model fit on both levels of analysis.

Three general types of suppression were discerned in the literature: classical, negative and reciprocal (mutual) (Conger, 1974). It seems that the above case is a good example of a reciprocal suppression when both parameters increase their value when suppressor is added. Hence, OCT indices do not yield classical suppression effect as was predicted, but reciprocal suppression effect. This is mainly due to a non-zero relation between measure of bias (suppressor) and math ability (criterion).

[100] Significance of regression weights change was calculated from the z distribution, while significance for R^2 was calculated from the F distribution. Percentage change was calculated using formula: $\frac{model\ 3*100\%}{model\ 1} - 100\%$.

Interpretation of Model 3 results is quite straightforward: in case of all three OCT scoring systems the measures of accuracy correlate positively with math ability, namely, the higher the self-reported familiarity with math terms the higher the objective mathematical abilities. In case of the IRT indices the measures of bias are interpreted as follows: the higher the claimed familiarity with foils the lower the objective math ability (Table 12). In case of the SDT indices the higher the predilection towards answering "yes" in math familiarity scale the higher the math ability as measured by the PISA test. It is noteworthy that foils claiming was related with low, not high math ability.

Only in case of the SDT-derived indices the Model 4 analysis confirms moderation effect, namely significant interaction between OCT accuracy and bias. It means that the effect of *d'* on math ability depends on *c* value. The below figure presents that the predictive validity of *d'* on math ability is highest when *c* is high. It is to say that relation between claiming familiarity with reals, but not foils, is more strongly related to math ability among participants that have a general tendency to use affirmative response categories in math familiarity scale. It is difficult to interpret this effect further on as to the best knowledge of the author interaction effects between SDT indices where not tested in the OCT literature, though their emergence is not shocking regarding the correlation between accuracy and bias measures. It is difficult to tell why the interaction term is not significant in case of IRT and common sense OCT indices, especially as common sense and SDT measures are very much related theoretically and statistically.

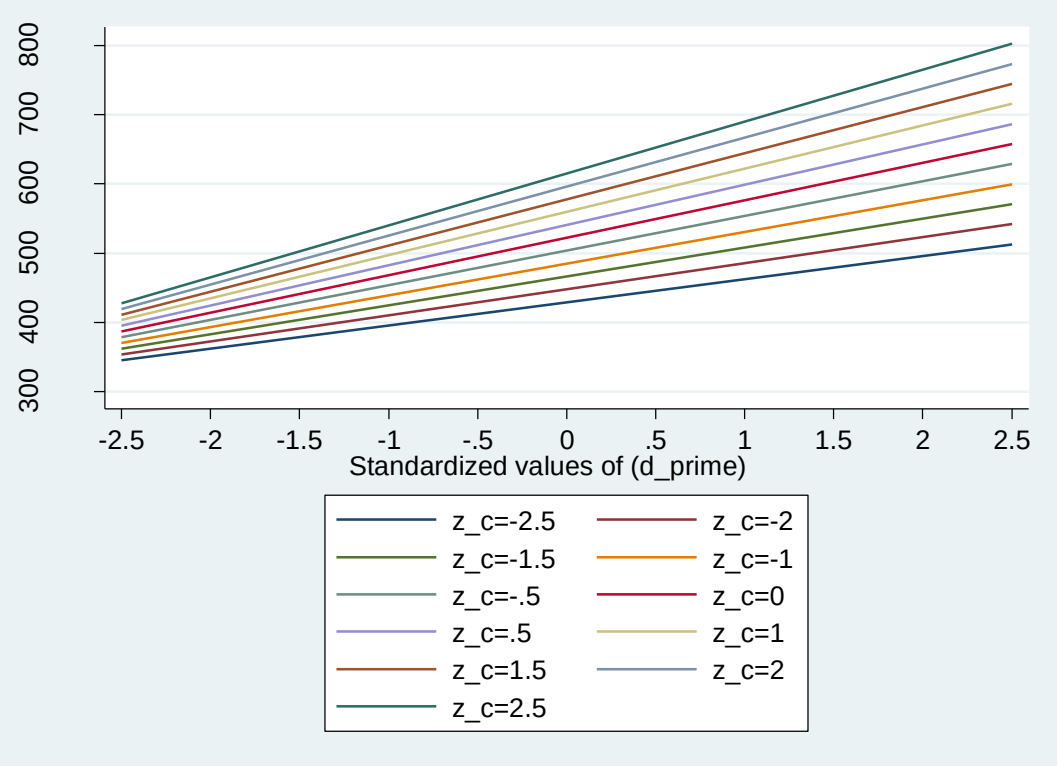

Figure 6. Margins plot for the interaction between d' and c and math ability predictions.

It is also very interesting that both IRT score for reals and for foils are more related to *c*, instead of measures of accuracy and bias being related to each other. It is warranted to suggest that the interpretation of *c* parameter as a measure of bias has to be rethought, especially in OCT versions where number of reals and foils is unbalanced. With the data at hand it can inferred from Table 11 that scores on both reals and foils depend primarily from the strategy that a given respondent adopts. In

this particular example the “liberal” strategy (claim familiarity/answer “yes”, use “4” and “5”) proved to be “efficient” strategy as it inevitably led to high scores on math familiarity scales. This was mainly driven by the unbalanced character of the OCT version used in PISA. With only three foils and 13 reals the use of this “liberal” strategy is bound to result with high math familiarity scores as the probability of answering to a real is more than four times higher than answering to a foil. It would be very interesting to verify how different scoring systems would behave in OCTs with different reals to foils proportions.

It is noteworthy, that *d’* correlated negatively and quite highly with IRT score for foils- it is a corroboration that both indices work as predicted. Moreover, IRT scores, in this form used for the first time to score an OCT, proved their worth yielding results clear to interpretation and consistent with predictions. They also noted the highest R^2 values and the regression weight for IRT accuracy index was the highest in magnitude, save for only the weight for *d’* from Model 4 where the added interaction term further enhanced this coefficient.

To sum up: OCT measures of bias indeed acted like a suppressor for the relation between math familiarity scale scores and math ability. However, reciprocal suppression pattern was yielded instead of the predicted classical suppression model. Nevertheless, this change does not have large interpretational consequences as all three suppression models are treated like continuum of sorts (Conger, 1974; Paulhus et al., 2004).

7.2 Overclaiming and memory bias- Hypotheses 2, 3 & 4

7.2.1 Method

Hypothesis 2, 3 and 4 consider related matters, hence they will be presented and discussed together.

In order to test these hypotheses pairwise correlation matrix as well as multilevel regression equations were analysed. The same statistical software and methodology was used as in case of Hypothesis 1 testing.

In order to test hypotheses 3 and 4 the following scales were employed:

- mathematical self-efficacy (st37)
- openness (st94)
- perseverance-giving up (st93)
- perseverance (st93)
- experience with pure (academical) math problems (st61)
- experience with applied math problems (st61)

The openness and perseverance scales used 5-point rating scales, whereas self-efficacy and experience scales used four response categories. All scales employed (in a sense) reversed scoring, where high scores (answering “4” or “5”) were related to low traits level, while low scores (embracing categories “1” and “2”) corresponded to high trait levels. To avoid confusion all scales were reversed so that high scores indicate high self-reported efficacy, openness, perseverance and experience and high propensity to give up easily (low perseverance). The last trait was expected to yield negative correlations with the remaining scales.

7.2.2 Results

The relation between OCT indices and math ability was in detail commented above and will not be presented here to avoid redundancy. However, they will be commented in the below 7.2.3 section, this time in the light of memory-related theories of overclaiming.

Moreover, the results for IA and IE are not shown as they are essentially redundant to the SDT indices (*d'* and *c*). What is more, the full correlation matrices are not presented either as a lot of information is negligible for the present hypothesis testing. All this information is available in Online Appendix C though.

Scale	**math ability**	**math familiarity (reals)**	**bias (foils)**	***d'***	**c reversed**
self-efficacy	0,65	0,54	ns	0,24	0,24
openness	0,29	0,30	ns	0,10	0,15
experience: applied tasks	0,16	0,21	0,08**	ns	0,17
experience: pure tasks	0,26	0,32	ns	0,16	0,13
perseverance	0,25	0,29	ns	0,09	0,16
giving-up	-0,20	-0,27	-0,08*	ns	-0,20

*Table 17. Pairwise correlations between math self-efficacy, openness, perseverance and math ability scored objectively (PISA test) and subjectively (math familiarity scale). Note: ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

The selected measures are positively related to measures of OCT accuracy, however the relation with IRT-scored reals is stronger than to *d'*-measured ability to discern reals from foils. However, the scale scores are not related negatively to measures of OCT bias: IRT-scored foils and *c*. The pattern of these relations is very similar between *d'* and *c* (small, positive correlations), whereas in case of IRT score for foils all coefficients are zero or close to it. However, to test the possibility that OCT bias acts as suppressor regarding the predictive validity of the above scales on math ability a regression analysis was run in order to check for each scale's unique predictive validity controlled for overclaiming tendency measured by IRT score for foils.

Parameter	**B**	***p***
math familiarity (reals)	22,77	***
bias (foils)	-11,10	***
self-efficacy	44,70	***
openness	-	ns
experience: applied tasks	-	ns
experience: pure tasks	5,51	*
perseverance	-	ns
giving-up	-	ns
R^2 level1	0,47	
R^2 level2	0,62	

*Table 18. Regression of math self-efficacy, openness, perseverance and math familiarity scale on math ability scored objectively (PISA test). Note: ns p> 0.05, *p<0.05, **p<0.01, *** p<0.001.*

The results in Table 18 show that from this set of self-report scales only math familiarity, overclaiming bias, experience with pure tasks and, to the largest extend, math self-efficacy have retained their predictive validities on math ability when controlled for other variables.

7.2.3 Discussion

Hypothesis 2 presumed that math ability will be related positively to OCT accuracy and negatively to OCT bias. The results presented in Tables 11-14 confirmed this claim regarding IRT-scored OCT indices, however the hypothesis was not corroborated regarding SDT (and "common sense") indices. Such patterns of results suggest that indeed math ability is related positively with both measure of math familiarity and measure of differentiation between reals and foils. However, regarding the traditionally-used bias measures, math ability correlates negatively with claiming familiarity with foils and correlates positively with a tendency towards "answering <<yes>> strategy" in math familiarity task (as measured by *c* and IE). However, this last relation may be mainly driven by the unbalanced character of the PISA 2012 OCT. With the data at hand it has to be said that the results obtained seem to confirm rather the metacognitive account than memory biased theory, as overclaiming seems to be an effect of lack of knowledge and failed metacognitive monitoring of the task rather than confirmatory bias stemming from large knowledge and overly liberal accepting of associations between real terms and foils.

Hypothesis 3 assumed that the self-report scales measuring other aspects of math ability (self-efficacy, experience) and cognitive functioning (perseverance and openness in problem-solving) will correlate positively with OCT accuracy and negatively with OCT bias. No such a pattern was observed in the data. The self-report measures correlated positively with OCT accuracy and positively with *c*, but did not correlate at all with the purest measure of overclaiming- IRT score for foils. This result further points that the so-called memory-based overclaiming theory was again disconfirmed. It seems that answering to foils is a distinct process, relatively unrelated to answering to other self-report scales, including reals from the same scale (but see subchapter 7.7 on more on this topic). This also seems to suggest that method variance between foils and other self-report scales is low.

When predicting math ability, both OCT accuracy and bias maintained their validity even controlled for other scales. Among all other scales analysed only self-efficacy yielded significant and substantially meaningful regression coefficient on math ability, other scales failed to reach statistical significance when controlled for OCT scores. When comparing Table 17 with Model 3 from Table 12 it becomes evident that self-efficacy has a large role in predicting math ability, but both OCT indices remain their predictive validity when controlled for this variable. In contrast, the predictive validity of all other scales analysed here was reduced to zero, when controlled for self-efficacy and OCT scores[101].

Hypothesis 4 predicted that math ability and self-reported mathematical skills will yield similar pattern of correlations with OCT scores. Comparing data from Tables 11 and 16 it has to be concluded that regarding SDT indices the pattern is indeed similar (positive correlations with both indices), but regarding the IRT indices it is not. Moreover, it has to be noted that correlations with objective math ability are higher than with subjective scales, including the relations with self-efficacy scale which yielded the highest relations to OCT accuracy and no relation to OCT bias. These results disconfirm Hypothesis 4 and also ask new questions about the role of self-perceived knowledge in overclaiming studied previously by Atir and colleagues (2015).

[101] This analysis was also performed using *d'* and *c*. The results were unchanged. This table is available in Online Appendix C.

This pattern of results warrants for a further (exploratory) check of the self-perceived knowledge and memory-based accounts of overclaiming by testing its domain-specificity *versus* domain-generality. To this end OCT indices were correlated with other abilities measured by the PISA test: reading and science.

Domain	math familiarity (reals)	bias (foils)	*d'*	c reversed
math ability	0,57	-0,11	0,34	0,16
reading ability	0,47	-0,15	0,33	0,07
science ability	0,47	-0,12	0,32	0,10

Table 19. Cross-domain relations of overclaiming. Note: zero-order correlations, all ps < 0.001.

Table 18 above indicates that OCT indices are related to a similar degree to each of the PISA 2012 domains. It is a clear evidence that OCT measures are related to a domain-general phenomenon instead of a domain-specific one.

It is noteworthy to compare the above results to the outcomes presented by Jerrim and co-authors (2019) who also analysed relations between self-efficacy, openness, perseverance and overclaiming. However, they did not use continuous OCT measures, employing instead a quartile split of OCT bias and comparing only the extreme groups (lowest OCT bias *versus* highest OCT bias). This dichotomised measure yielded significant relations with self-efficacy, openness and perseverance measures, with participants scoring high on OCT bias noting higher, more socially desirable, scores on these measures. Replicating these analyses in this work yielded similar results in case of openness and a subscale of perseverance (giving up). However, there is ample evidence that such splitting of continuous data is not recommended as it increases both Type II error (through loss of power) and Type I error (MacCallum, Zhang, Preacher & Rucker, 2002; McClelland, Lynch Jr., Irwin, Spiller & Fitzsimons, 2015). Moreover, when controlling for *c,* such quartile results ceased to be significant[102] suggesting that any observed differences are spurious, driven probably by stylistic factors (e.g. ERS). These ideas will be further addressed at the end of this chapter.

To sum up: the above analyses suggest to support the metacognitive monitoring over memory-biased, overgeneralised associations account as a possible mechanism of overclaiming (cf. Paulhus & Dubois, 2014). They also show that overclaiming is related to domain-general phenomena, as it is not related to other self-report measures of math ability and it is not limited to one cognitive domain only. This also disconfirms memory bias hypothesis and suggests alternative explanations.

7.3 Overclaiming and motivated response biases- Hypotheses 5, 6 & 7

7.3.1 Method

If not commented elsewise all the methodological details were similar as in the above analyses.

In order to test hypotheses 5 and 6 the following scales were employed:

- instrumental motivation to learn math (st29)
- interest in math (st29)

[102] Consult Online Appendix C for details.

- social norms regarding learning math: parents (st35)
- social norms regarding learning math: friends (st35)
- math anxiety (st42)
- math self-concept (st42)
- math learning work ethic (st46)
- intention to learn math in future (st48)
- math-related behaviours (st49)

All these scales were used as coding various aspects of math importance, desirability and positivity (emotions related to learning math).

In order to test hypothesis 7 the following scales were used:

- locus of control (attribution of success in learning mathematics) (st43)
- attribution of failure in school context (st44)
- attitude towards school utility (learning outcomes) (st88)
- attitude towards hard work at school (st89)
- success control in school (st91)

Save the intention to learn math in future items all other scales used four response categories. All scales employed (in a sense) reversed scoring, where high scores (answering “4”) were related to low traits level, while low scores (embracing categories “1”) corresponded to high trait levels. To avoid confusion all scales were reversed so that high scores indicate high trait levels. High trait levels in case of the attribution of failure scale meant attributing the failure to internal causes. The intention to learn math in future scale (st48) used dichotomous scoring, where low scores (“1) coded intention to learn math, whereas high scores (“2) coded intention to learn other subjects. The scale coding was reversed and due to its response format the scale was scored by IRT 2PLM model instead of GRM.

7.3.2 Results

The results for Hypotheses 5 and 6 are summed up in the table below:

Scale	**math ability**	**math familiarity (reals)**	**bias (foils)**	***d'***	**c reversed**
instrumental motivation	0,29	0,30	0,09**	ns	0,23
math interest	0,26	0,29	0,11**	ns	0,24
math anxiety	-0,54	-0,44	ns	-0,16	-0,23
math self-concept	0,55	0,44	0,09**	0,12	0,23
work ethic	0,14	0,29	0,08*	ns	0,27
learning behaviour	0,13	0,24	0,12	ns	0,22
future intentions	0,25	0,18	0,07*	ns	0,16
norms: parents	0,14	0,18	0,08**	ns	0,15
norms: friends	-0,18	ns	0,08*	-0,09	0,08

*Table 20. Pairwise correlations between math interest, importance and effort and math ability, scored objectively (PISA test) and subjectively (math familiarity scale). Note: ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

The above results brought some confirmation for Hypothesis 5: OCT bias is indeed related positively with math desirability, as evidence by almost all scales presented in Table 19. The sole exception was math anxiety scale that did not correlate with IRT foils score, thus disconfirming Hypothesis 6.

The results for Hypothesis 7 are presented below:

Scale	**math ability**	**math familiarity (reals)**	**bias (foils)**	***d'***	**c reversed**
control of school success	0,12	0,13	ns	ns	0,09**
attribution of failure	0,19	0,22	0,09*	ns	0,19
locus of control	0,28	0,28	ns	ns	0,17
school utility (learning outcomes)	ns	0,11	0,10**	-0,07*	0,17
school utility (learning activities)	0,08**	0,17	ns	0,05	0,18

*Table 21. Pairwise correlations between school control, locus of control and school-related attitudes and math ability, scored objectively (PISA test) and subjectively (math familiarity scale). Note: ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

Most of the presented scales yielded very small or even non-significant correlations with OCT indices. The strongest pairwise relations were established with attributing failure towards external causes (instead of internal) which correlated negatively with math ability and positively with bias, as measured by IRT index for foils. This last result, albeit tiny in size, confirmed somewhat predictions from Hypothesis 7 regarding OCT bias. Also in line with this hypothesis is the general correlation between control of school success and OCT accuracy measures.

7.3.3 Discussion

Hypothesis 5 was confirmed to some extent as it was evidenced that item desirability was indeed related with overclaiming. However, the correlations with foils index were not high and some of them were significant only at the 0.05 level. This seems to suggest that self-enhancement tendencies indeed played some role in overclaiming familiarity with mathematical concepts. This result point to the utility of including an OCT in the questionnaire and also that it is advisable to measure individual item desirability as it is predictive of response biases.

Lack of relation between math anxiety index and OCT bias disconfirmed Hypothesis 6. However, this result may stem from the specificity of the math anxiety scale used in the PISA 2012 questionnaire. The PISA version of this scale has good psychometric qualities but the content covered is very restricted in comparison to other, more established math anxiety scales, e.g. Abbreviated Math Anxiety Scale (AMAS; Cipora et al., 2015; 2018). It was assumed that math anxiety was negatively related with desirability, but it is possible that the negative correlations between these scales should not be interpreted in this way ("I am afraid of math, so I do not think it is important") and that alternative explanations should be employed (e.g. relation due to mutual correlations with math ability).

It is interesting to observe that the scales did not correlate with *d'* but all of them correlated with *c* and math familiarity. This points to special care that needs to be paid for OCT scoring system and its implications as it may bring non-trivial interpretations for overclaiming nomological network. Again it

was evidenced that *c* cannot be treated as a pure measure of bias, to which goal an IRT-scored foils are much better (see similar thoughts in Atir et al., 2015). It is also interesting that *c*, as an indicator of a general tendency towards answering "yes" to items is more strongly related to the wide spectrum of scales than *d'*. It turns out that the ability to discern reals from foils is based on processes that are very loosely related to high scores in other math-related scales. Thus, it is possible that *d'* is a context-specific index that should not be used in every analysis. Moreover, it is to be determined whether the relation between *c* and other scales is driven by stylistic (e.g. ERS or ARS), motivated (SDR, S-E) or substantial (math ability) factors.

Hypothesis 7 was confirmed only to a very limited degree as only the failure attribution scale correlated with OCT bias. It is unclear why of the three scales measuring (in theory) similar concepts in the area of locus of control[103] only this one was related with OCT bias. The internal locus of control was previously related with overly positive biases, mainly unrealistic control (Hoorens & Buunk, 1993). However, other research linked external locus of control with academic cheating and dishonesty (Coleman & Mahaffey, 2000). The results presented by Griffith and colleagues (2006) yielded a very complicated pattern of dependencies between locus of control and faking, as in some domains it was internal locus of control predicting faking, whereas in others it was external that was correlated with faking.

Regarding the scales analysed in Table 20 it is important to notice that some of them had low psychometric quality (see Online Appendix B) and was based on a minimal number of items- three. These factors lower the utility of these scales and potentially also had impact on the pattern of results presented in Table 20.

An interesting result obtained is the correlation between utility of school outcomes with OCT bias- this may be an indication of SDR as in general it is not socially correct for a 15-years-old student to openly admit that school has given him/her nothing. Hence, this positive relation may be a measure of "politically correct" answering to the items from the school outcomes scale. This interpretation is further corroborated by correlations yielded by this scale with SDT indices- the relation with *d'* is negative, whereas correlation with *c* is positive. This translates to an interpretation that high scores in this scale are related to higher claiming familiarity with foils, less differentiation between foils and reals and general tendency to use more positive response categories. All this points to an SDR distortion in this scale. Moreover, this pattern may be a key to distinguish distorted scales- they should probably jointly correlate with *d'* (negatively) and *c* (positively) as correlations with only one of the SDT indices may not be diagnostic (Paulhus et al., 2003).

7.4 Overclaiming and motivated response biases- Hypotheses 8, 9 & 10

7.4.1 Method

The general methodological approach related to testing these hypotheses is similar to this elected in the above subchapters. However, in this part also data from school questionnaire will be analysed and data from school principals will be confronted with students' answers.

The following scales from the student questionnaire will be used in subsequent analyses:

[103] Locus of control is a psychological trait indicating the degree to which people attribute life outcomes to internal *versus* external causes. An example of an internal locus of control is a student attributing a high score on a math exam to her hard work and good preparation. A student with external locus of control would attribute such success to luck or test easiness, while failure would be attributed e.g. to teacher's unfairness, bad luck, test difficulty, etc., namely causes beyond student's control.

- truancy (st08, st09, st115)
- school disciplinary climate (st81)
- teacher-student relations (st86)
- sense of belonging to school (st87)

All of this scales used a four-point response scales. In case of truancy items high scores indicated high levels of truancy (skipping classes, coming late to school, etc.). In case of the remaining scales high scores indicated low levels of traits, hence they were all rescaled so that higher scores indicated better school climate, teacher-student relations and higher sense of belonging.

The school questionnaire scales employed in this part were:

- use of assessments in school policy (sc18)
- use of assessment results in accountability process (sc19)
- school disciplinary climate (sc22)
- parental pressure on high academic achievement (sc24)

The sc18 and sc19 scales used dichotomous scoring, school disciplinary inventory employed four-category response scale, while parental pressure consisted of just one item with three possible response categories. All scales were coded and scaled so that high score indicate high trait level. The three truancy items were joint to form one scale. The sc19 and sc24 did comprised less than three items, hence they were used as single items indicators.

To calculate discrepancy between principals' and students' views on school climate two methods were used: a) calculating difference between two scores, b) employing a suppression model.

7.4.2. Results

The pairwise correlations analysis brought the following results:

Scale	**math ability**	**math familiarity (reals)**	**bias (foils)**	***d'***	**c reversed**
truancy	-0,16	-0,17	ns	-0,09	-0,08
teacher-student relations	-0,06*	ns	0,08**	-0,08*	0,14
sense of belonging	ns	0,10**	0,07*	ns	0,13
disciplinary climate (students)	ns	0,15	ns	ns	0,10
assessments in school policy	ns	ns	ns	ns	ns
school climate (principals)	ns	ns	ns	ns	ns
accountability: public	0,11	0,11*	ns	ns	ns
accountability: authority	ns	ns	ns	ns	ns
parental pressure	ns	ns	ns	ns	ns

*Table 22. Pairwise correlations between school climate and school accountability and school-related attitudes and math ability, scored objectively (PISA test) and subjectively (math familiarity scale). Note: ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

Higher truancy correlated negatively with math ability, both objectively and subjectively scored, but did not correlate with OCT bias. This measure did correlate with SDT indices but yielded a slightly unexpected pattern of negative correlations with both *d'* and *c.* It means that students reporting higher truancy had lower ability to differentiate reals from foils and also had lower tendency to answer "yes" in OCT scale.

The three scales related to school discipline yield only minimal correlations to OCT indices. Interestingly, two of these scales correlated positively with OCT bias. Especially the pattern of the teacher-student relations scale correlations was interesting as it showed positive correlation with OCT bias, but negative with OCT accuracy (and math ability).

The scales from the principals' questionnaire failed to correlate with anything. The sole exception were small correlations yielded between using achievement data publicly and math ability. It may be indicative of some motivational effect or of using this accountability procedure only by certain type of schools.

Neither of the analyses using discrepancy index between principals and students views on school discipline yielded significant results. The details can be found in Online Appendix C.

7.4.3 Discussion

Hypothesis 8 assumed that participants reporting less school-related problems (e.g. truancy, disciplinary troubles) would also yield higher OCT bias. This hypothesis was partially confirmed as participants reporting higher school-belonging and better teacher-student relations also had higher OCT bias and lower OCT accuracy (in case of the teacher-students relations scale). This may be indicative of an SDR distortion of their responses. Similar results and similar explanation was offered by Jerrim et al. (2019), thus further corroborating this line of interpretation.

Similarly as in the paper by Jerrim and co-authors (2019) truancy was not related to OCT bias. This is certainly a surprising result, as truancy is, in theory, a non-desirable behaviour which is not expected of "good" students. Yet, these scales seem not distorted by any kind of self-enhancing tendencies. It is a question for future research projects why is it so. Perhaps students do not see these behaviours as "bad" and, consequently, item desirability of such scales is low and does not ignite SDR tendencies? Comparing self-reported truancy with more objective data, e.g. from school administration records, would also help to disentangle this puzzling example.

It is also important to notice that most of the relations in this part is small, which may point to non-substantial interpretations, e.g. method variance causing these tiny correlations (cf. Khorramdel & von Davier, 2014). Some of the variables presented here was also based on single items from the school questionnaire so at best they can be treated as slight indications of future research directions, not firm evidence.

Hypothesis 9 was not confirmed as difference between principals' and students' views on school disciplinary climate did not correlate with any of the OCT measures. It is concluded that this idea of research seems a good idea of a thorough investigation, preferably in a design enabling application of the multi-trait, multi-method model (MTMM) (see Podsakoff et al., 2003). In this work using such models was not possible due to lack of adequate data (only one similar questionnaire rated by both students and principals).

Hypothesis 10 was not confirmed as no information on school policy correlated with OCT measures. Only use of achievement data in school accountability correlated positively with math ability but this effect is beyond the scope of this study and will not be discussed further on.

7.5 Overclaiming and careless responding- Hypotheses 11 & 11a

7.5.1 Method

Pairwise correlations, regression interaction terms and analysis of cut-off points were employed in order to test relations between various C/IER measures and OCT scores analysis.

Among the C/IER measures proposed in literature a wide plethora of *post-hoc* indices was calculated (remedial internal methods in the terminology of the subchapter 4.5). However, only one external remedy measure was possible to employ, simply because only one such set of items was included in the PISA 2012 student questionnaire.

Altogether the following indices were calculated and applied in analyses:

- Mahalanobis distance
- dr* distance
- psychometric synonyms
- even-odd consistency/Cattell's sabotage index
- multiple fixed individualised chance score (MFIC)
- IRT personal fit statistics for polytomous data: G, U3, *lz*
- intra-individual response variability (irv)
- long-string measure
- average string length
- self-reported effort invested in solving PISA tasks (clcuse301, 302, deffort)

Outlier indices

Mahalanobis distance and dr* distance measures represent examples of outlier method analysis, popular in regression diagnostics. Mahalanobis distance returns a value indicating distance from data centre in a multivariate distribution. This measure serves for finding outlier, aberrant participants in the dataset. Dr* is a generalisation of Mahalanobis distance for ordinal self-report data recently proposed by Mansolf and Reise (2018) to serve as a person-fit measure.

Other person fit statistics, G, U3 and *lz*, are extensions of IRT person fit measures for polytomous data. They serve in order to identify and flag aberrant response vectors in the data analysed.

G is also known as Guttman errors and it is a number of response categories embraced that violate certain assumptions of an underpinning IRT model. Each step (response category) on a ordinal scale has its estimated difficulty parameter, thus it is assumed that a respondent has to have a given trait level in order to pass a given step. For example, on a math efficacy using four-category rating scale (0-to-3, with "3" denoting maximum trait level) item "I can solve integrals" is almost certainly more difficult (on average in the whole sample) than item "I can calculate how much cheaper a TV would be after a 30% discount". Hence, respondent embracing category "3" in the "integrals" item and selecting only category "1" in the "TV" item has yielded an unexpected pattern, as she embraced a very high (difficult) response in an item of high difficulty but failed to reach such a high response on a much easier item. Thus, this respondent would have committed a Guttman error. The precise number of Guttman errors is calculated as a number of easier steps missed (categories not selected) when more difficult item steps were passed (categories selected). The higher the number of Guttman errors, the more aberrant the response vector is assumed (Emons, 2008).

U3 is based on a very similar idea as Guttman errors and also denotes the number of easy steps missed, while more difficult steps were embraced. The index is scaled to a 0-1 range, with values closer to "1" indicating higher misfit (Emons, 2008).

Lz is a person-fit statistics based on a value of standardised log-likelihood for an observed vector of responses. The index has its critical values set on the basis of distributional assumptions at a given level of significance (Emons, 2009). Small values of this statistic are indicative of aberrant responses.

Individual consistency

Psychometric synonyms index resides on finding pairs of items that are on average correlated with each other above certain criterium (e.g. at least 0.5). When such pairs are found in the whole sample a within-person correlation is calculated for all such pairs to form a value of the psychometric synonyms index. In this work two thresholds were used: 0.5 and 0.4.

The even-odd measure *alias* Cattell's (sabotage) index, known also as intra-individual reliability, is a within-person correlation between scores calculated for even and odd items separately. The two measures are conceptually very similar but have been calculated using different methods: even-odd index just calculated the within-person correlation between the two scores, whereas the Cattell's index was based on residuals from a model where even items' score was regressed on odd items' score (cf. Fronczyk, 2014).

Invariability methods

Intra-individual response variability (irv) is a within-person standard deviation calculated for a given set of data. Low values of this index may be indicative of straightlining, while very high values may suggest (pseudo)random responding (Dunn, Heggestad, Shanock & Theilgard, 2018; Marjanovic, Holden, Struthers, Cribbie & Greenglass, 2015).

Long-string measure is a number of identical responses in a row (uninterrupted string of identical responses). If few such strings are present in a given participant's response vector the highest number of them is given as this measure's value. Average string length indicates average number of identical responses in a row in a given vector.

MFIC is an index derived from calculating the proportion of a given response category used in the scale. The index is higher when some response categories are used more often than others.

A consummate source of information on C/IER and its indices can be found in the article by Meade and Craig (2012). Other sources offering sumptuous information on the matter are e.g. Curran (2016), Fronczyk (2014), Huang, Curran, Keeney, Poposki and DeShon (2012) or Johnson (2005).

External remedy methods

The PISA 2012 student questionnaire included two items measuring self-reported students' engagement in solving PISA tasks. In one of these items students assessed their effort put in PISA on the 1-to-10 scale (with "10" indicating the highest effort), while in the other they assessed their effort if the PISA was a high-stakes exam. The PISA 2012 database also includes a difference between these two items as a measure of attitude towards PISA test (how motivated a student was in comparison to a "real" test) (see more details on this matter in OECD, 2019).

Measure of respondents' fatigue

Due to the PISA rotational design participants solve item booklets that contain partly overlapping content but often in different order. The design for the PISA 2012 student questionnaire was

schematically shown in Table 5. Math familiarity scale was included in forms A and C of background questionnaire booklets, where in form A it was placed at the very end of questionnaire, while in form C it was placed in the middle. It was assumed that participants in form C could display less C/IER due to lower fatigue. If OCT scores are affected by respondents' fatigue they should also differ between the two forms.

C/IER indices cut-offs

Apart from using the C/IER indices in their continuous form they were also employed to flag aberrant response vectors in the math familiarity scale. In order to flag such respondents cut-offs logic was used, namely participants exceeding certain values on a given index were flagged as potential outliers. The person-fit measures were calculated using the "PerFit" R package (Tendeiro, 2018). In case of the remaining indices the criterion of two and half standard deviations was used: participants with values on a given index above or below its value (depending on the side of distribution diagnostic for aberrant responses) were indicated as potential careless respondents. The criterion-related validity of the math familiarity scale as well as values of OCT measures were compared between the outlying (careless) and "regular" (attentive) participants. Similar approach to flagging C/IER outliers was recommended e.g. by Reise et al. (2016).

Software

The R package "careless" (Yentes & Wilhelm, 2018) was used to calculate most of the C/IER indices, save the person-fit measures that were estimated using the "PerFit" R package (Tendeiro, 2018) and MFIC and even-odd measures which were brought by calculations performed in Stata on the basis of equations provided by Fronczyk (2014). The dr* index was calculated in R using the code provided by the authors of the measure (Mansolf & Reise, 2018)[104].

[104] I would like to thank Paweł Grygiel and Grzegorz Humenny who obtained the code from the authors and taught me how to use it.

7.5.2 Results

Pairwise correlations between C/IER indices and OCT measures and math ability are presented in the table below:

C/IER index	**math ability**	**math familiarity (reals)**	**bias (foils)**	***d'***	**c reversed**
Mahalanobis distance	-0,24	-0,35	-0,06	-0,06	-0,28
irv (intra-individual response variability)	0,19	0,12	-0,46	0,44	-0,41
psychometric synonyms 0.5	0,38	0,45	-0,20	0,33	ns
psychometric synonyms 0.4	0,39	0,43	-0,32	0,48	-0,11
even-odd correlation	-0,26	-0,28	0,25	-0,33	0,06
long-string	-0,12	-0,07	0,21	-0,15	0,13
average string	-0,15	ns	0,30	-0,23	0,23
G person-fit statistic	-0,27	-0,42	0,08	-0,22	-0,16
G^r	-0,30	-0,47	ns	-0,16	-0,25
Lz person-fit statistic	0,17	0,12	-0,39	0,35	-0,29
lz^r	0,09	0,08	-0,17	0,13	-0,13
U3 person-fit statistic	-0,23	-0,30	0,20	-0,27	ns
$U3^r$	-0,17	-0,22	0,10	-0,13	ns
dr*	0,23	0,30	-0,21	0,26	-0,10
Cattell's index	ns	ns	ns	ns	ns
MFIC	0,29	0,24	-0,44	0,52	-0,26
effort PISA	ns	ns	ns	-0,05*	0,05*
difference effort	ns	ns	ns	ns	ns

Table 23. Zero-order correlations between C/IER indices, OCT measures and math ability (objectively and subjectively measured).

*Note: r- index calculated only basing on reals, foils excluded; ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

Most of the C/IER indices yielded significant zero-order correlations with math ability and OCT measures. These relations go in the predicted direction, e.g. *lz* correlates positively with bias which is perfectly predictable taking into consideration that low values of this statistic are indicative of aberrant responses, while U3 correlates negatively with OCT bias and positively with OCT accuracy which is also predictable as high values of this index indicate outlying patterns.

The highest correlations with OCT measures were yielded by MFIC, *lz*, irv, average string length and psychometric synonyms (threshold 0.4), whereas the lowest correlations were noted by G and Mahalanobis distance. Cattell's index failed to correlate with anything. Problems with this measure where communicated earlier on (e.g. O'Dell, 1971) so most likely more research is needed on its calculation and theoretical foundation to claim its utility.

Also self-reported effort and difference between self-reported effort on PISA test and high-stakes test did not correlate with any other measures. The minimal correlations yielded between self-reported effort and OCT measures were not significant when correction for multiple comparisons was applied. This result points to further research on self-reported measures of effort as their utility still seems questionable.

After analysing zero-order correlations, moderator roles of C/IER indices were tested in regression equations for reduction and moderator effects on OCT bias measures. Cattell's index and self-reported test effort were not included in these analyses. Moreover, due to high correlations between certain indices[105] only some of the indices were included in regressions. Only cases with no missing values in the math familiarity scale vector were included in the below analyses (cf. Reise et al., 2016). **Standardised versions of the C/IER indices were used in order to ease the interpretation of regression coefficients.** Original (raw scale) measures differ greatly in their scales, ranges and standard deviations.

Math ability was regressed on math familiarity (IRT index for reals), overclaiming (IRT index for foils) and a given C/IER index. Two models were compared: 1) model with only OCT measures included, 2) model with C/IER index and interaction term C/IER index*OCT bias added. R^2 change was omitted as there was no model in which it would exceed 0.02. For all calculations the -mi estimate- Stata package was used.

[105] Such correlation matrix is available in Online Appendix C.

C/IER index	C/IER index B	OCT bias B change (in %)	interaction term B
Mahalanobis distance	-5,40**	2,29%	ns
irv (intra-individual response variability)	5,60*	-24,42%	ns
psychometric synonyms 0.5	10,68***	-42,10%	-3,86*
psychometric synonyms 0.4	12,75***	-85,50%	ns
even-odd	-4,89*	-23,81%	ns
long-string	ns	-12,77%	ns
average string	ns	-16,50%	ns
G person fit statistic	-5,95**	-3,28%	5,23**
G^r	-5,70**	-1,25%	4,49**
lz person fit statistic	ns	-16,00%	ns
lz^r	ns	-3,46%	ns
U3 person fit statistic	-6,75**	-9,02%	3,72*
$U3^r$	-5,31*	-3,75%	ns
dr*	4,15*	-9,23%	ns
MFIC	9,94***	-44,52%	ns

Table 24. C/IER indices as moderators of OCT bias.

*Note: r- index calculated only basing on reals, foils excluded; ns p> 0.05, *p<0.05, **p<0.01, *** p<0.001.*

Adding C/IER indices to the above-mentioned regression equation resulted in reduction of the OCT bias coefficient. The coefficients for the added C/IER indices were not of particularly high value. This seems to suggest that most of the C/IER indices acted like a redundant predictor, adding which to regression equation results in reduced weighted validity of other predictors. Such interpretation is corroborated by insignificant changes of R^2 parameter.

The above regressions were also calculated for SDT indices. The results were essentially the same, hence their presentation is omitted here.

Interestingly, some of the C/IER indices yielded significant interaction terms with OCT bias (G, Gr, U3, psychometric synonyms 0.5). Marginal effects for the interaction between OCT bias and G (Guttman errors value) is presented in the figure below:

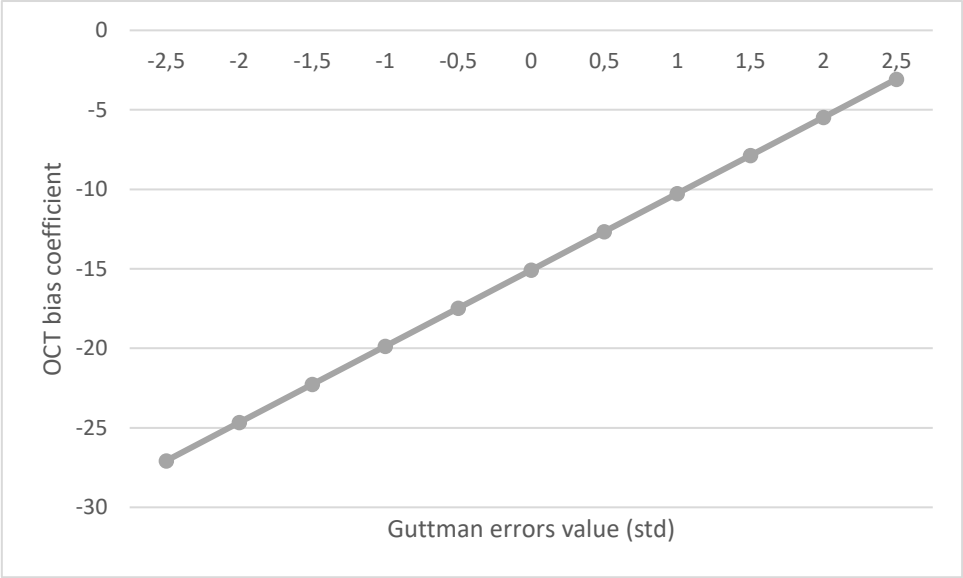

Figure 7. Marginal effects for OCT bias Guttman errors interaction.*

It is evident from the figure above that the higher the value of G (indicating increasingly aberrant pattern of responses) the lower the predictive validity (represented by regression weight) of OCT bias on math ability. Such pattern of results would suggest that C/IER is indeed related partially to overclaiming of mathematical terms. Some of the high values of OCT bias may be derived from careless responding instead of other mechanisms, typically linked with overclaiming, e.g. positivity bias or memory bias.

The above analyses showed that C/IER indices share some portion of variance with OCT bias but do not act as suppressors in regression equations. Most of the C/IER indices do not moderate the OCT bias's relation with math ability neither, with a notable exception of G, U3 and psychometric synonyms. It was observed that in the area of distribution of C/IER index indicating aberrant patterns OCT bias loses its predictive validity on math ability. However, most of the C/IER indices were not developed for treating them as continuous variables but were rather built up in order to flag abnormal patterns of responses and foreclose them from further analyses.

Such approach was undertaken in next analysis where OCT measures predictive validity was compared for normal and aberrant response patterns. Even-odd, dr* and MFIC measures were dropped from these analyses, because only a handful of participants exceeded the limit of two and a half standard deviation (or even two standard deviations). The results are displayed in the table below:

C/IER index	outlier		normal		% flagged	tail	Cut-off criterion
	OCT accuracy	OCT bias	OCT accuracy	OCT bias			
Mahalanobis distance	28,27*	ns	49,45	-15,02	3,09%	right	2.5 SD
irv (intra-individual response variability)	36,67	ns	50,87	-13,77	2,81%	left	2.5 SD
psychometric synonyms 0.5	46,38	ns	48,71	-15,02	4,41%	left	2 SD
psychometric synonyms 0.4	48,37	ns	48,54	-14,02	3,23%	left	2 SD
long-string	38,61	ns	51,26	-13,49	3,99%	right	2.5 SD
average string	40,20	ns	50,85	-13,56	3,33%	right	2.5 SD
G person fit statistic	ns	ns	50,13	-15,41	2,46%	right	2.5 SD
G person fit statistic	39,07	ns	49,46	-15,01	11,00%	right	individual fit cut-off value
G^r person fit statistic	ns	ns	49,84	-15,25	2,53%	right	2.5 SD
lz person fit statistic	ns	ns	49,93	-15,08	2,81%	left	2.5 SD
lz person fit statistic	30,88	ns	49,75	-14,69	8,96%	left	individual fit cut-off value
lz^r person fit statistic	25,98*	ns	49,48	-14,63	2,53%	left	2.5 SD
U3 person fit statistic	ns	ns	51,01	-15,27	2,64%	right	2.5 SD
U3 person fit statistic	33,29	ns	50,02	-14,73	7,95%	right	individual fit cut-off value
$U3^r$ person fit statistic	17,37*	ns	50,85	-14,89	2,53%	right	2.5 SD

Table 25. Regression weights of OCT accuracy and OCT bias on math ability in aberrant (careless) and normal subgroups.

*Note: r-statistics calculated only for reals, foils excluded. ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

Cut-offs for psychometric synonyms were set as values below -2 standard deviation as there were too few cases below -2.5 standard deviation to estimate the needed model. For G, u3 and *lz* apart from calculating the values for standard deviation cut-offs also cut-off values based on these statistics' distribution were calculated (Tendeiro, 2018).

All statistics brought largely comparable results as the subgroup flagged as outlying yielded much less predictive validity of OCT measures on math ability. In some cases (psychometric synonyms, irv, long-

string and perfit measures with distributional/individual fit cut-offs) the regression coefficient for OCT accuracy was not drastically different between the two subgroups, but in most of the cases these coefficients were seriously underestimated in the aberrant group. In all cases the flagged group had severely distorted OCT bias coefficients that were close to zero in most of the indices. It is interesting, that person-fit measures with cut-offs set on the basis of individual fit values, despite flagging a much larger percent of the sample than other measures, yielded similar results but with important difference: the OCT accuracy coefficients in this subgroup were much closer to the coefficients obtained from the normal subsample, but OCT bias coefficients were still aberrant in this case. This points to a conclusion that much larger variance of OCT bias was related to C/IER than in case of OCT accuracy and that this is especially true in the 5-10% of the most aberrant response patterns.

As an exploration of this conclusion the correlation between OCT accuracy and bias was calculated in the deciles of *lz* statistic (each decile counted 288 respondents). The results of this exploration are presented in the figure below:

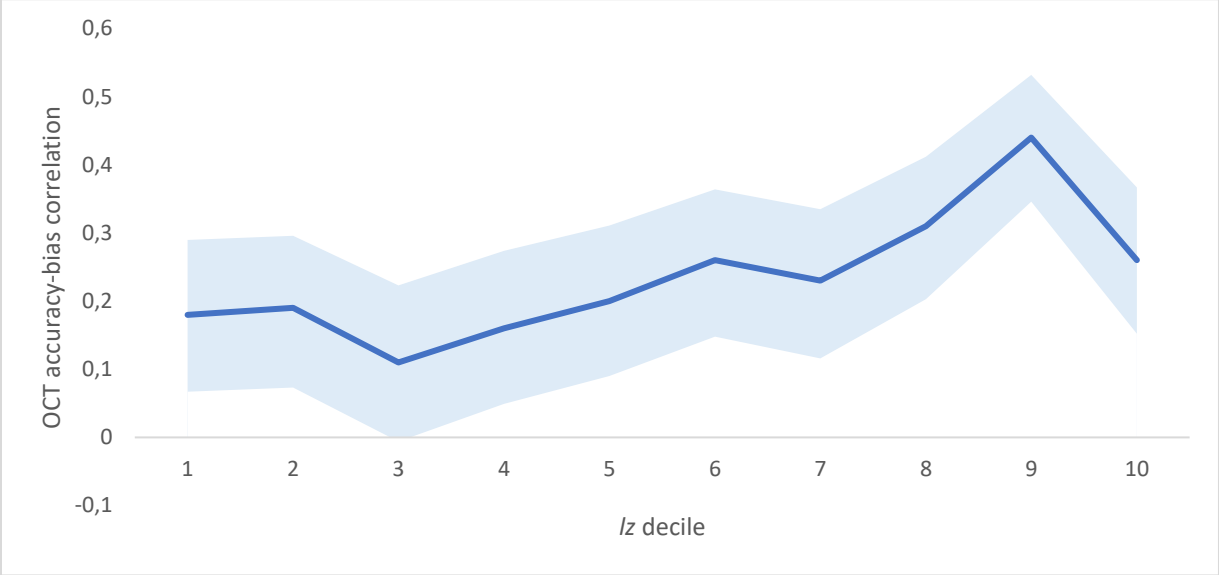

Figure 8. Correlation between OCT accuracy and OCT bias depending on the lz decile.

The correlation is a bit lower in the left tail of the *lz* distribution but the confidence intervals[106] (95%, light blue in the figure above) are largely overlapping, indicating that the differences between deciles are not huge and mostly not significant. However, there is a certain trend towards lower relation between these scores in the area of the *lz* statistic that may denote C/IER.

The above analyses for SDT indices of OCT are essentially leading to the same conclusions, thus they are not presented in order to avoid redundancy.

Respondents fatigue and overclaiming

Multilevel regression with the fatigue condition as 0-1 variable was added in order to account for the PISA rotational design effect on OCT measures:

[106] Confidence intervals were calculated using user-written Stata package -ci2- (Seed, 2002).

IRT OCT indices	*B*	*p*
math familiarity (reals)	54,88	***
OCT bias (foils)	-17,08	***
fatigue	-	ns
fatigue*math familiarity	-9,40	**
fatigue*OCT bias	-	ns
SDT OCT indices	***B***	***p***
d’	60,99	***
c reversed	42,10	***
fatigue	-	ns
d’*c reversed	10,24	***
fatigue*d’	-11,44	**
fatigue*c reversed	-	ns
d’*c reversed*fatigue	-	ns

*Table 26. Respondents fatigue as induced by PISA rotational design and OCT measures. Note: B-regression weights, PISA scale where 1SD=100; ns p> 0.05, *p<0.05, **p<0.01, *** p<0.001.*

Fatigue was not significant as a main effect, nevertheless, it turned out to yield a significant interaction term with measures of OCT accuracy: math familiarity IRT index and *d’*. In case of both indices fatigue condition decreased OCT accuracy predictive validity by around 20%. Surprisingly, fatigue has no effect on OCT bias measures. Fatigue also indicated significant zero-order correlations with C/IER indices that went into predicted direction- the fatigue condition was related to more C/IER as compared to the non-fatigue condition. The coefficients were of very similar size $r\sim|0.09|$, $p<0.0001$, but failed to yield any significant main or interaction terms with C/IER indices in a series of multilevel full factorial regression equations of math ability on OCT measures, C/IER indices and fatigue condition[107].

7.5.3 Discussion

The C/IER indices turned out to be a useful addition for modelling relations between OCT measures and math ability. Most of the indices yielded zero-order correlations with these variables into the predicted direction. However, few of them failed to produce significant correlations with any of the measures (e.g. Cattell’s index), others brought in relations in counter-logical direction (e.g. even-odd).

Self-reported measures of test effort did not yield any significant correlations with OCT measures, however, it is rather a result that points to the limited utility of such measures (especially one item ones) and not that OCT bias cannot be driven by C/IER (cf. Jerrim et al., 2019).

In most of the cases the careless responding indices failed to act as suppressor or moderator variables in multilevel regression equations. When introduced into regressions they only reduced the regression weight of OCT bias but failed to bring in unique variance. Thus, they can be assumed as largely redundant predictors in these equations (Conger, 1974). However, some of them produced significant and substantially meaningful interactions with OCT bias measure- bias’s predictive validity was seriously decreased in the areas of the C/IER index distribution indicative of distorted responding. Thus, it was assumed that indeed overclaiming is partially related to careless responding and that C/IER could be validly considered as mechanism creating OCT scores. However, experimental designs are

[107] These results are omitted due to redundancy.

indispensable to test such theories and verify their verity. For now it is impossible to conclude anything more than a trace of such relation has been found and it is ready for further scrutiny.

Moreover, also the C/IER indices themselves require further research as some of them seem of dubious utility or yield surprising, counter-predictive results (if not to say non-sensical). Interesting option for further investigation is how different C/IER indices should be combined in order to bring maximum efficiency of identifying aberrant patterns in the data (cf. Meade & Craig, 2012).

Fatigue, as induced by the PISA rotational design, correlated with C/IER indices (more C/IER in the fatigue condition) and moderated predictive validity of OCT accuracy. Surprisingly, it was not related to OCT bias scores. Moreover, it did not produce significant interaction terms with C/IER indices in regression equations. More research is needed on this topic, especially with the use of experimental studies with fatigue manipulated both between and within conditions.

To sum up: both Hypothesis 11 and 11a were confirmed. Moreover, interesting results were identified. Careless responding is a potential mechanism responsible for some part of the OCT bias variance. Furthermore, C/IER indices moderated correlation between reals and foils in OCT. This is an interesting track for future investigations in this area.

7.6 Overclaiming and response styles- Hypothesis 12

7.6.1 Method

Among the available methods of RS modelling the IRTree method was chosen due to its well-established position and a growing body of evidence of valid results it yielded (Khorramdel & Von Davier, 2014; Khorramdel et al., 2019; Plieninger, 2020a).

The IRTree method, introduced by Böckenholt (2012), resides on recoding the observed responses into a set of binary variables, called “pseudo-items” or “pseudo-responses” (Plieninger, 2020a). This is done in order to disentangle different variance sources underlying the observed responses. In this approach it will be assumed that there are three such sources: the target trait, extreme response style (ERS) and midpoint response style (MRS). ERS is a tendency to use only extreme response categories (e.g. “1” and “5”), whereas MRS is a predilection towards middle response categories (e.g. “3”). The pseudo-items construction and coding process follows logic depicted below:

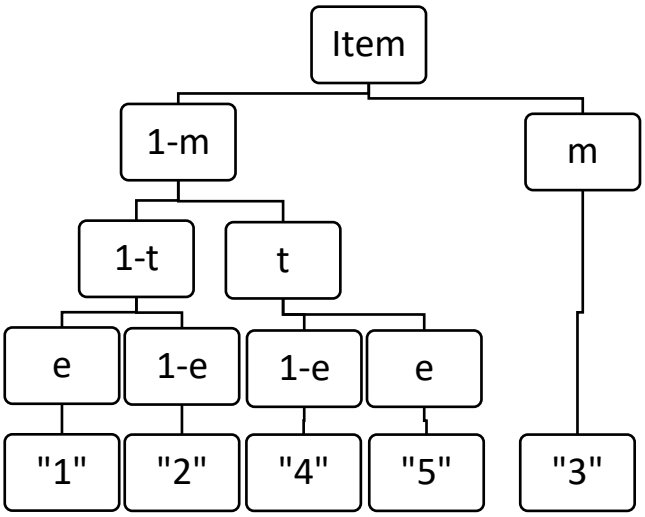

Figure 9. Schematic representation of the pseudo-items coding system.

It is assumed that at first a respondent decides whether she is embracing a middle response category (m choice) or not. If "3" is chosen than the whole decision process ends and respondent passes to another item. If "3" is not chosen (1-m choice), then respondent decides whether she possess trait measured by the item (e.g. agrees with the item, claims familiarity, etc.). This is a t choice (have trait) *versus* 1-t choice (do not have trait). In the next step the respondent has to take yet another decision: whether she decides to use the extreme response category (e choice) or not (1-e choice). If an extreme option is chosen than the respondent embraces "1" or "5" (depending on the t choice made). If an extreme option is not chosen then the respondent embraces "2" or "4" (again depending on the previous choice t *versus* 1-t). This decision tree is used to map respondents' choices in the so-called mapping matrix. The above coding is only one of the many proposed in the field (Falk & Cai, 2016; Plieninger, 2020a), but the most researched one so far (Khorramdel & von Davier, 2014).

The mapping matrix resulting from the above coding is depicted below:

Item response	**Pseudo-item coding**		
x	X_{trait}	$X_{extreme}$	$X_{midpoint}$
5	1	1	0
4	1	0	0
3	-	-	1
2	0	0	0
1	0	1	0

Table 27. Pseudo-items mapping matrix.

The R "mirt" package (Chalmers, 2012) was chosen as a software used to calculate multidimensional IRT models and "ItemResponseTrees"[108] wrapper was used to model mapping matrices for the pseudo-items (Plieninger, 2020b). Two models were estimated for further comparisons: a) model with ERS, MRS and trait responses (three-dimensional model that assumes ERS and MRS presence in the data), b) one-dimensional GRM with no RS modelled (it is assumed that there are no RS in the data). Both models were estimated by the expectation-maximisation (EM) algorithm and by the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010). The results were literally identical as the estimations for RS factor scores correlated on the level 0.9996. Thus the two estimations will not be compared here and the estimates from the EM algorithm will be reported. The data was weighted by applying the final student weights. Three estimates for each student were made: a) estimate for trait "ability", b) estimate for ERS "ability, c) estimate for MRS "ability". In case of both models the trait responses were modelled by the GRM, while in the three-dimensional model the RS factors were estimated using PCM, as it was assumed that every item contributes to the RS dimension equally (Plieninger, 2020a).

In order to deepen the analyses and explore differences between in-subscale and across-subscale RS (Khorramdel & von Davier, 2014) three sets of RS scores were calculated: a) where both reals and foils from the math familiarity scale were used to calculate them, b) where only reals were used, and, c) where only foils were used.

[108] I would like to express my sincere thanks for Hansjoerg Plieninger for introducing me to this very useful package.

The estimated RS results will be correlated with OCT scores in order to check if any relations between these processes exist. The Stata package -repest- was used to calculate these correlations.

7.6.2 Results

Comparing fit between the one-dimensional (no RS in the data) and the three-dimensional model (ERS and MRS in the data apparat from the construct of interest) is a necessary first step in any RS analysis (Pokropek et al., in preparation):

Model	AIC	BIC	log Likelihood	parameters	observations	dimensions
1-d (no RS)	9197648,98	9198478,99	-4598744,49	69	2881	1
3-d (ERS, MRS, trait)	*9075094,56*	*9075810,44*	*-4537478,28*	80	2881	3

Table 28. Goodness of fit for models modelling response style presence in the data. Note: RS- response style, ERS- extreme response style, MRS- midpoint response style.

Analysis of the goodness of fit measures showed that the three-dimensional (3-d) model fitted the data better than the one-dimensional (1-d) model, hence confirming that RS are present in the data, as it was expected for the PISA 2012 dataset (Khorramdel et al., 2017).

The correlations between the scores for three pseudo-items are presented in the table below:

Pseudo-item score	**trait**	**ERS**	**MRS**
trait	1	-	-
ERS	0,18	1	-
MRS	-0,30	-0,47	1

Pseudo-item score: foils	**trait**	**ERS**	**MRS**
trait	1	-	-
ERS	-0,54	1	-
MRS	0,34	-0,59	1

Pseudo-item scores: reals	**trait**	**ERS**	**MRS**
trait	1	-	-
ERS	0,35	1	-
MRS	-0,47	-0,48	1

Table 29. Correlations between pseudo-item scores. Note: ERS- extreme response style, MRS- midpoint response style.

These pattern of results is also expected, as typically trait of interest (construct variance) correlates negatively with MRS and positively with ERS, moreover, the two RS measures are most often negatively associated (Plieninger, 2020a; Pokropek et al., in preparation). What is interesting in this particular example is that the latter correlation is rather small in size, as in some examples MRS and ERS correlated much more negatively with each other (e.g. Pokropek et al., in preparation: ~-0.80), though it is also has to be admitted that any evidence in the field of RS is rather thin (e.g. Plieninger & Meiser, 2014 also obtained Ers-MRS correlation in the range of ~0.50). Furthermore, trait and ERS seem to be slightly less related to each other than it was previously reported (e.g. Plieninger & Meiser, 2014; Pokropek et al., in preparation, obtained this correlation in the range of 0.4), though the matter was never systematically scrutinised as so far. Probably the specificity of the math familiarity questionnaire (foils presence chiefly) plays some role in shaping these relations.

Correlations between RS scores for foils yielded some interesting differences when compared to such correlations for indicators estimated solely from the reals responses: in case of foils, trait and ERS scores correlated negatively and trait and MRS correlated positively, whereas in case of reals these relations were reversed.

Nevertheless, the main point of interest is to gauge the relation between RS and OCT measures. These correlations are presented herein:

Pseudo-item score	**math ability**	**math familiarity (reals)**	**bias (foils)**	***d'***	***c* reversed**
trait	0,30	0,68	0,42	-0,07**	0,79
ERS	0,37	0,53	-0,17	0,39	ns
MRS	-0,27	-0,43	0,08	-0,27	-0,10
trait: reals	0,39	0,75	0,30	0,11	0,70
ERS: reals	0,40	0,60	-0,08**	0,32	0,13
MRS: reals	-0,32	-0,52	ns	-0,24	-0,22
trait: foils	-0,07**	0,21	0,73	-0,65	0,78
ERS: foils	0,15	0,07**	-0,51	0,56	-0,46
MRS: foils	-0,13	-0,07	0,37	-0,42	0,32

*Table 30. Correlations between RS scores and OCT measures. Note: ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

ERS yielded positive relation to math ability (both measured objectively and subjectively), whereas MRS was negatively associated to it, which is a typical pattern for zero-order correlations of RS scores (Plieninger & Meiser, 2014).

Interestingly the construct indicator for foils correlated negatively with OCT accuracy as measured by *d'* but positively when measured by IRT score for reals- this result suggests that there is a common trait underpinning responses to both kinds of these items. ERS for foils correlated negatively with OCT bias, both SDT and IRT scored. This result is surprising at the first glance but it is probably driven by a large proportion of "1" responses for foils in the sample (such response loads ERS but correlates negatively with both measures of OCT bias; see also Figure 5 for further comparisons). From the comparison of

MRS scores correlations for reals and foils it can be suggested that MRS correlates positively with a trait variance when it denotes an above-average response, but correlates negatively when it denotes a below-average response. A substantial in size negative correlation between *d'* and trait score for foils would suggest that in this OCT version avoiding false alarms was a much more important loading of OCT accuracy (SDT scored) than scoring hits. Table 29 also informs that answering to foils was marginally related to math ability, with claiming familiarity with foils being a negative predictor of math attainment.

7.6.3 Discussion

In contrary to what was predicted by Hypothesis 12 ERS was related positively only to OCT accuracy, but not OCT bias. In fact, this indicator was related negatively to OCT bias, regardless whether it was calculated solely from reals, foils or from both type of items together. On the other hand, MRS calculated from foils only yielded negative relations to OCT accuracy and positive to OCT bias. It seems that a lot of foils variance stems from midpoint responses. It is up to discussion whether these stem from a genuine conviction of some familiarity (substantial response) or whether they are a result of avoidance, careless responding, puzzlement or confusion caused by "strange" items (stylistic response). Nevertheless, there is no evidence for overclaiming as a side-effect of ERS, however, in this OCT version it was MRS that was related to claiming familiarity with non-existent math concepts.

7.7 Latent structure of the PISA 2012 overclaiming scale- Hypothesis 13

7.7.1 Method

Structural validity is one of the aspects of construct validity (Messick, 1995). Modelling the OCT's latent structure precisely would help to answer questions related to the mechanisms underlying the observable OCT scores. Performing such an analysis makes possible to further inform an important question: whether participants use same, similar or dissimilar mechanisms when answering to reals *versus* foils? The first assumption would mean that only one factor should emerge or, alternatively, two, strictly correlated factors should form. However, using dissimilar processes to answer reals and foils should lead to an emergence of two (or more) relatively unrelated factors. It is also interesting whether such factors would yield any cross-loadings, indicating that certain foils might be considered as reals by participants and *vice versa*: some reals could be treated as foils (cf. Leite & Cooper, 2010).

Hence, in order to determine the PISA 2012 OCT structure the following models were compared: a) one-factor model with all indicators loading on it, b) two-factor model, with one factor for reals, one for foils, c) higher order solutions. The best fitting model would be considered the most applicable for the PISA 2012 math familiarity scale for the Polish sample.

To perform all the necessary calculations Mplus 8.2 software and Stata 14.2/SE standard procedures were used. Data were clustered at the school level and both student and school weights were applied.

7.7.2 Results

Before fitting confirmatory factor analysis (CFA) an exploratory factor analysis (EFA) was applied in order to scout the math familiarity internal structure. EFA yielded a two-factor oblique structure of the data[109]. The table presenting factor loadings is presented below:

Item	Factor1	Factor2	Uniqueness
st62q01		0,35	0,73
st62q02	0,65		0,55
st62q03		0,44	0,67
st62q04		0,45	0,64
st62q06		0,63	0,56
st62q07		0,35	0,73
st62q08		0,88	0,34
st62q09	0,63		0,53
st62q10	0,90		0,31
st62q11		0,79	0,49
st62q12	0,91		0,28
st62q13		0,57	0,70
st62q15	0,57		0,61
st62q16		0,51	0,75
st62q17	0,80		0,42
st62q19	0,61		0,60

Table 31. Factor loadings matrix from EFA. Note: Loadings < 0.30 are suppressed.

The solution shows a somewhat surprising image: two correlated (r=0.43) factors emerged, but the classification of items to factors does not resemble anything we would have expected for basing on the theoretical scale structure. First of all, reals and foils do not form separate factors, instead, all three foils are clustered within Factor 2 but accompanied by a bunch of reals. Moreover, few items (st62q01, st62q07, st62q16) have really high uniqueness, indicating that they do not share much variance with other items of the scale.

The above analysis indicates that the scales' structure may be more complicated that thought previously. The models displayed below were estimated using the Weighted Least Square Mean and Variance (WLSMV) estimator hence no direct comparison of the models, e.g. through AIC/BIC statistics, is possible. However, the comparison of the goodness of fit statistics displayed below gives a good impression onto which model accounts for the observed variance best[110]:

[109] Maximum likelihood estimation used, promax (power=3) rotation applied to the factor matrix.

[110] In all models the SRMR statistic equaled to 0.176 on the between level (school-level) as no additional school-level variables were added.

Model	1	2	3	4
	one-factor	two-factor	bifactor-REALS/FOILS	bifactor-EASY/HARD
χ^2	7863,762	6372,519	5522,442	1149,684
df	208	206	204	192
p	***	***	***	***
RMSEA	0,113	0,102	0,095	0,042
CFI	0,696	0,755	0,789	0,962
TLI	0,65	0,715	0,752	0,953
SRMR-within	0,175	0,154	0,139	0,052

*Table 32. Goodness of fit statistics- comparison between the estimated CFA models for math familiarity scale. Note: *** p < 0.001.*

The comparison between the single order models clearly favoured a non-unidimensional solution. In the two-factor model mathematical ability and responding to foils and factors were correlated with each other (~0.60)[111]. Subsequently, a series of analysis with higher-order models was performed. A bifactor model with one general factor, one specific for foils (plus item st62q08 which loaded more on the foil factor than math factor) and one specific for reals was estimated. This model did not yield any spectacular increase in goodness of fit. Subsequently, a bifactor model with one general factor and two specific factors: one for easy[112] items, one for hard items, was estimated. This was the first model which yielded goodness of fit statistics close to the acceptable values[113] (cf. Hu & Bentler, 1999). It was assumed that this was a satisfying version of the math familiarity scale model as it informed the basic research questions[114].

[111]Moreover, neither of the models fit the data well (RMSEA ~ 0.150, CFI/TLI~0.750, SRMR>0.100). Compare with similar observations in Pokropek (2014).

[112] Easiness or difficultness was assessed basing on the items' means in the sample. Items of means above 3 were classified to the "easy" specific factor.

[113] Of course the "accepted values" were achieved only regarding the descriptive statistics of goodness of fit. The χ^2 goodness of fit test was still significant, yielding insufficient similarity of the model-implied covariance matrix to the empirical covariance matrix. However, the seriousness of this misfit is still debated in the field: some proponents claim that this test is impractical due to its shortcomings (e.g. dependence on the sample size and model complexity) (Schermelleh-Engel, Moosbrugger & Muller, 2003; Tanaka, 1987), others reject it completely, especially in large sample (N > 1000) studies (Barret, 2007). However, there is a group of researchers who strongly opposes such treatment of the χ^2 goodness of fit test. For example, the team centred around Leslie Hayduk (e.g. Hayduk et al., 2007) points to a great importance of this test and orders strict treatment of it- a failed χ^2 test means that a model has failed. Such models should be discussed in categories "why did it fail" and not as if they were "OK anyway". Moreover, Hayduk et al. (2007) contend that a failed χ^2 could be indicative of serious casual model misspecification. However, the working consensus in the field seems to reserve from the two above propositions and the general advice is to treat this test only as one of many sources of information on the goodness of fit (e.g. MacCallum, 2003). Jöreskog & Sörbom (1993) advocate dropping the dichotomous approach towards the test (fit/no fit) and treat it as a descriptive statistic with large values indicating bad fit and lower values- better fit. See also the discussion in Konarski (2009, pp. 328-335). Taking this latest advice it is evident that Model 4 from Table 32 displays much better fit than other verified models as evidenced by a large drop of the χ^2 value.

[114] The correlated-factors model was also specified and estimated but it did not converge with the use of the WLSMV estimator. Hence, it was decided not to include its results in these analyses. The model probably did not converge due to only two level-1 factors.

These analyses indicated that items were most likely were sorted between the factors mostly on their difficulty, not OCT type (foil vs. real). Especially items st62q08 and st62q04 were eager to "change sides": the former was more related to the bias factor, whereas the latter to the math ability factor (see also subchapter 6.2.2 and 6.2.3). Definitely, the PISA 2012 OCT was not a unidimensional scale, however, reals and foils did not form stable separate factors. It seems that to many of the respondents reals and foils looked much alike and there is no trace of any specific variance part that could be attributed to responding to the foils only.

The below table presents the individual-level factor loadings and item R^2 (item residual variances are just 1-R^2) of the math familiarity scale items[115]:

Item	general factor	specific factor: easy	specific factor: hard	*R2*
st62q01	0,55	-	0,24	0,36
st62q02	0,74	0,24	-	0,61
st62q03	0,57	-	0,33	0,43
st62q04	0,60	-	0,34	0,48
st62q06	0,49	-	0,54	0,53
st62q07	0,56	-	0,25	0,37
st62q08	0,28	-	0,80	0,71
st62q09	0,71	0,30	-	0,59
st62q10	0,61	0,66	-	0,80
st62q11	0,06	-	0,89	0,79
st62q12	0,62	0,70	-	0,87
st62q13	0,22	-	0,59	0,39
st62q15	0,65	0,31	-	0,52
st62q16	0,29	-	0,46	0,30
st62q17	0,63	0,56	-	0,72
st62q19	0,57	0,44	-	0,52

Table 33. Factor loadings and R^2 for the bifactor solution of the math familiarity scale- individual level.

The above table reveals that despite modelling each item by two factors some of them still have a rather low R^2 (e.g. st62q01, st62q13- a foil, or st62q16), meaning that the model still does not account for their variance very well. Moreover, it is evident that some of the foils load quite well on the general factor (deemed as math ability), e.g. st62q04.

As the model accounted for the clustered nature of the data also the between-level item statistics were generated. They are displayed in the subsequent table:

[115] More information on the models fitted in this section is available in the Online Appendix E.

Item	general factor	specific factor	*R2*
st62q01	0,39	hard	0,15
st62q02	0,94	easy	0,89
st62q03	0,55	hard	0,30
st62q04	0,40	hard	0,16
st62q06	0,24	hard	0,06
st62q07	0,57	hard	0,32
st62q08	0,21	hard	0,04
st62q09	0,80	easy	0,64
st62q10	0,92	easy	0,84
st62q11	ns	hard	0,14
st62q12	0,87	easy	0,75
st62q13	ns	hard	0,12
st62q15	0,85	easy	0,73
st62q16	0,35	hard	0,12
st62q17	0,93	easy	0,86
st62q19	0,71	easy	0,51

Table 34. Factor loadings and R^2 for the bifactor solution of the math familiarity scale- school level. Note: ns- loading was not statistically significant.

It is striking that the school-level factor did not account for any variance in foils st62q11 and st62q13 (non-significant factor loadings). Moreover, the model estimated accounted in general very well for the school-level variance of the easy items, but not for the hard items. It seems that school-level is not related to foils responses (see also subchapters 7.8 and 7.9 on this topic), on the other hand, the between-school variance of the easy items is well accounted for by the model. However, the between-school variance of the hard items[116] remained largely unaccounted for by the model. Answering to this question is beyond the scope of the present work.

Bifactor ancillary statistical indices[117]

In order to enhance model's interpretation the so-called bifactor ancillary statistical indices were calculated (Rodriguez, Reise & Haviland, 2016). To this end automatic scripts embedded in the Dueber's (2017) calculator were used.

[116] It is possible that items' difficulty is further convoluted with items wording complexity as some of the items are dyadic, consisted of two words. These dyadic items, e.g. "complex number", "linear equation", etc., could have yielded substantial common variance with foils due to the sole fact of their complex structure. It is to be reminded that all foils constituted of two terms, just like as these items did. It is thus possible, that complex, dyadic items were for some reason more difficult (more suspicious?, more difficult to process linguistically?) for participants which led to their low embracement and clustered them together with foils in one factor. It is also conceivable that the reals' content, pertaining to algebra or geometry, could also generate some confounding variance, although the analysis of specific factors seems to promote items' difficulty and then complexity, not represented area of mathematics, as main reason of items' sorting between the factors (see also OECD 2014a).

[117] These indices are new developments and to the best knowledge of the author their properties were not thoroughly tested in the multilevel contexts. However, such indices are indeed used and reported in the multilevel bifactor models (e.g. Wang et al., 2018), hence it was decided to use them here too.

Factor/Index	ECV (1)	ECV (2)	ω/ ω_S	ω_H/ ω_{HS}	Relative ω	H	FD
General Factor	0,522	0,522	0,932	0,642	0,689	0,886	0,905
Specific Factor 1- Easy items	0,186	0,362	0,929	0,310	0,334	0,729	0,855
Specific Factor 2- Hard items	0,292	0,600	0,877	0,527	0,601	0,878	0,933

Table 35. Factor-level ancillary bifactor indices. The measures are on the 0-1 scale.

The explained common variance (ECV 1 & 2) indicates what is the proportion of common variance explained by the general factor. Values above 0.80 are considered indicative of an essential scale's unidimensionality. In case of the PISA 2012 overclaiming scale the value of ECV for general factor (0.522) clearly indicates that this scale should not be modelled by a one-dimensional model. The ECV 1 values show that both specific factors account for significant portions of the common variance (0.186 and 0.292 for easy items and hard/SDR items respectively). Moreover, the ECV 2 values indicate that the specific factors account for an even more significant portion of the common variance in case of the items loading on them- values of 0.362 and 0.600 are almost double in comparison to the ECV 1 ones (see Stucky & Edelen, 2015 for more information on ECV 1 & 2).

Omega indices for both general and specific factors, representing internal reliability under the model fitted to the data, indicate good reliability of all three factors specified in the model as they are well above the 0.80 threshold (Reise, Bonifay & Haviland, 2013). However, omega hierarchical, denoting what proportion of variance in total (factor) scores can be attributed to the general factor, yielded a much lower value of 0.642. The comparison between the values of ω and ω_H in this model shows that around 29% of the total variance can be attributed to the specific factors (ω-ω_H) and around 7% to random error (1-ω ; see Rodriguez et al., 2016 for more on these indices). Omega HS index reflects the reliability of a specific factor score after controlling for the variance attributable to the general factor (Rodriguez et al., 2016). The values of 0.310 and 0.527 show that scores of the specific factors are not very reliable and using them as outright variables e.g. in an SEM model is not warranted in this case. It is mainly due to the fact that little variance resembles after accounting for the general factor (see DeMars, 2013 for more on score interpretation in bifactor models).

The H values, which denote construct replicability, namely how well is a given latent variable represented by the set of observable items and how well is this latent variable expected to replicate in other studies (Hancock & Mueller, 2001), indicate that all three factors can be quite successfully replicated in future studies, as they all exceeded the 0.70 threshold advocated by the authors of the measures (and the general factor and the specific factor for hard/SDR items even exceeded the more demanding 0.80 threshold). The factor determinacy (FD) statistic denotes the correlation between factor scores and the factors, with values above the 0.90 threshold indicating that respective factor scores can be used as variables on their own (e.g. in a regression or an SEM model). Again values for the general factor and the specific factor 2 indicate better "trustworthiness" of these scores, but the value for the specific factor 1 (easy items) is not far from the 0.90 threshold (Rodriguez et al., 2016).

The ancillary bifactor statistics comprise also item-level statistics that are displayed in the table below:

Item	IECV	ARPB
st62q01	0,84	0,004
st62q02	0,91	0,039
st62q03	0,75	0,047
st62q04	0,76	0,048
st62q06	0,45	0,292
st62q07	0,83	0,009
st62q08	0,11	1,286
st62q09	0,85	0,001
st62q10	0,46	0,348
st62q11	0,01	8,117
st62q12	0,44	0,403
st62q13	0,12	1,118
st62q15	0,82	0,02
st62q16	0,28	0,528
st62q17	0,56	0,232
st62q19	0,63	0,165
Average ARPB	-	0,791

Table 36. Item-level ancillary bifactor indices. Note: IECV- item explained common variance, ARPB- absolute relative parameter bias.

The item explained common variance (IECV) values act as a measurement of item-level unidimensionality with values above 0.80-0.85 indicating sufficient representation of the general dimension by the item variance. As it can be seen in the above table only a small subset of the PISA 2012 overclaiming scale would be successfully modelled by a unidimensional model. Items such as st62q01, st62q62q02 or st62q09 are represented enough by the general factor, however, items like st62q08 or st62q16 are poorly accounted for by this dimension. It is interesting that among the three foils two are almost not represented by the general factor at all (st62q11 and q13), whereas the remaining st62q04 is fairly well accounted for with the IECV value of 0.76.

The absolute relative parameter bias (ARPB) denotes differences between item loadings in the unidimensional solution and in the bifactor model (Dueber, 2017). Large discrepancies are indicative of unsuitability of the unidimensional model to account for a given item set. In this case the averaged ARPB shows that the differences between the two models are too large to accept unidimensional model as an accurate model for the PISA 2012 overclaiming scale (values around 0.10-0.15 are considered maximal for the average ARPB if a given model is to be accepted as a unidimensional one; Muthen, Kaplan & Hollis, 1987).

7.7.3 Discussion

Hypothesis 13 was somewhat confirmed as math familiarity scale indeed better fitted to the multi-factor solution than to a unidimensional structure. Moreover, the two emerging factors were correlated, as predicted in the hypothesis. However, this structure did not fit to the observed data (cf. Pokropek, 2014) as a closer examination of the factor structure revealed that the assumption that foils are a pure measure of bias, whereas reals constitute a pure measure of math ability is not justified. The EFA and CFA analyses revealed items loading on different factor that they were supposed to do or even an entirely different structure than expected. On this level of knowledge about OCT it is impossible to definitively discern what mechanism stood behind the observed pattern.

Moreover, it is possible that the math familiarity scale's latent structure was best modelled by a bifactor solution due to its superior ability to account for spurious variance in the data (Reise et al., 2016). Furthermore, there is also a certain arbitrariness in specifying the lower order factors, that were constructed in a more exploratory than confirmatory process and may not hold for a different set of items[118]. However, it is worthy to point out that both bifactor solutions tested in the model, the one with specific factors for reals and foils as well as the one with specific factors for hard and easy items, were theoretically justified (Goecke et al., 2020; Pokropek, 2014), nonetheless, neither of them was empirically tested to date. From the analysis conducted above it could be inferred that item difficulty resulted to be much more important than items' ontic status. It is for further analyses to discern what item characteristics exactly play key role in self-report of skills. It is noteworthy, that such studies would be informative not only for the OCT research field but also for any survey using self-report of skills as these characteristic interact with the processes used to respond to foils and reals alike.

Other higher-order models, e.g. random-intercept EFA models (Aichholzer, 2014; 2015), seem a promising avenue of future studies on OCT structural validity. Such endeavours would also benefit from the enhanced design of the OCT itself, e.g. only three foils is an absolute minimum of indicators for any kind of EFA/SEM analysis and this number should probably be higher in the next versions of the method. Moreover, both reals and foils should be also carefully piloted and matched on their difficulty and other characteristics such as word length, composition and similarity to other concepts (cf. Goecke et al., 2020). Linguistic studies on word recognition offer here an ample source of knowledge from which future designs of OCT should not hesitate to dip up. Moreover, such studies should also account for the crossed characteristics of items in the PISA 2012 math familiarity scale were all foils were consisted of two words (dyadic items), whereas reals consisted of both dyadic and simple items. Furthermore, the scales used in ILSAs are notorious for their latent structure problems as they are usually not developed to have simple structure but to cover the substantial content (van Dijk, Datema, Welten & van de Vijver, 2009).

Another important direction for future is to clarify the now tangled interpretation of the correlation between reals and foils. The above analyses show that it probably has a substance (e.g. math ability) substrate but what other variance causes this relation is a matter for future research to pinpoint. However, the relation between objective ability and reals-foils correlation (interactional model with ability as a moderator of the reals-foils relation) was rejected in a recent study of Goecke and collaborators (2020) where general intelligence (*gc*) failed to moderate such correlation. It is up for further studies to establish the nature of this relation and its contextual covariates.

[118] See Urban, Szigeti, Kokoneyi & Demetrovics (2014) for a study where method factor specification was much eased by the previously gathered evidence.

Moreover, it would be interesting to verify how math ability level influences the observed relationships. Whether foils are just difficult reals for both high-achieving and low-achieving students? Interestingly it was suggested before, although with the use of different items, that existent, but just very difficult items may also work as foils in an overclaiming measure (Steger et al., 2020). It seems that this effect has been somewhat replicated here as foils and difficult reals formed one specific factor. Such a pattern of results points that more knowledge should be gathered on the implications of various foils construction rules (see also Franzen & Mader, 2019 and Goecke et al., 2020 on a similar topic). An experimental study that would more systematically compare results for different types of foils (cf. Hargittai, 2005) and contrast them with responses to very difficult reals (e.g. concepts that could be hardly expected to be known in a given sample) would potentially bring valuable data on the processes observed in the above analysis. Others important item characteristics, e.g. word frequency, word length, similarity to existing terms, content area, etc. should also be tested in a similar way (cf. Goecke et al., 2020 for some ideas on this matter).

The ancillary bifactor indices offered additional support for the bifactor structure, indicating that unidimensional model was not appropriate for the data modelled and that the general factor was a reliable and replicable representation of the general variance in the data. However, both specific factors, but especially the specific factor for the easy items, did not reach that high levels of reliability as the main factor did, indicating that their scores (residual scores of the common variance controlled for the general factor; DeMars, 2013) should be treated with more care than the general factor scores when used as dependent or independent variables in other models, e.g. regression equations or SEMs.

7.8 School-level overclaiming correlates- Hypotheses 14 & 15

7.8.1 Method

The below analyses are planned in order to verify hypotheses relating overclaiming to school-levels of traits related to math ability, pressure to math-related achievements and rule breaking. Variables were selected from the PISA 2012 dataset on the basis of previous, individual-level analyses presented in subchapters 7.2 to 7.4. Specifically, self-efficacy, variables enlisted in subchapter 7.3.1 (save control-related variables that failed to produce significant effects on individual-level) and variables enlisted in subchapter 7.4.1.

The school-level variables were achieved by averaging the individual scores on a given variables. The final student weights were applied and standard errors were clustered at the school level (cf. Jerrim et al., 2019).

Pairwise zero-order correlation coefficients were used in order to explore dependencies between these variables and OCT measures and multilevel regression equations were employed to analyse first-order associations. In order to analyse the individual- and school-level effects separately the method of contextual effects was used (Christ et al., 2014). Identifying such effects would mean that school-levels of certain variables (e.g. math anxiety) are related to the modelled dependent variables additionally and irrespectively to the individual-level values of these variables (Raudenbusch & Bryk, 2002).

7.8.2 Results

The results of zero-order correlations of variables related to math achievement with OCT measures are summed up in the table below:

School-level mean of:	**math ability**	**math familiarity (reals)**	**bias (foils)**	***d'***	**c reversed**
norms: parents	ns	ns	ns	ns	ns
norms: friends	ns	ns	ns	ns	ns
self-efficacy	0,35	0,27	-0,06*	0,19	0,05*
work ethic	ns	0,07*	ns	ns	0,06**
learning behaviour	ns	ns	0,06*	ns	0,07**
future intentions	0,09*	0,09**	ns	ns	ns
math interest	ns	0,06*	ns	ns	ns
instrumental motivation	ns	0,09**	ns	ns	0,05**
math anxiety	-0,26	-0,20**	ns	-0,14	ns
math self-concept	0,23**	0,17**	ns	0,11*	ns

*Table 37. Correlations of school-level means of math achievement-related variables and pressure towards mathematic attainment with math ability and OCT measures. Note: ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

All the variables analysed yielded non-significant or minimal relations with OCT bias indices. Interestingly, math work ethic, learning behaviour and instrumental motivation were related to OCT accuracy measured by IRT index for reals and were associated to *c*, but were not related to objectively measured math ability. Such pattern may be indicative of spurious relations, probably driven by ERS or ARS as suggested by positive relation with *c* reversed (indicating higher tendency for affirmative responding). It is also worthy to comment that OCT accuracy measured by IRT index of reals was related to a higher number of variables than OCT accuracy measured by *d'*. Regarding the very small size of these relations it can be concluded that they are of probably spurious nature. *D'* seems to be a measure that is less prone to such biases than IRT accuracy index.

The results for school-level rule breaking are shown below:

School-level mean of:	**math ability**	**math familiarity (reals)**	**bias (foils)**	**d'**	**c reversed**
disciplinary climate (students-reported)	0,10*	0,12	ns	ns	0,07
truancy	ns	-0,05*	ns	ns	-0,06**
sense of belonging	ns	ns	0,04*	ns	0,07**
teacher-student relations	ns	ns	ns	-0,05*	0,06**

*Table 38. Correlations between school-level rule violation, math ability and OCT measures. Note: ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

From the indices analysed only sense of belonging to school was related to OCT bias measured by IRT score for foils: the higher the school belonging, the higher the bias. However, it is probably only a spurious relation, due to its small size and simultaneous relation with *c*. Among other results school-level truancy produced a negative relation to OCT accuracy. However, this relation is also very small in size.

7.8.3 Discussion

Hypothesis 14 was not confirmed as evidence for the relation between school competitiveness and pressure for math achievement with overclaiming was very scarce. Hence, it is another small and indirect evidence against considering overclaiming as an effect of positivity bias.

Hypothesis 15 was not confirmed as school-level rule violation was related to OCT in an only very limited way. Hence, there is no firm evidence for an association between prevalence of undesirable, immoral behaviours and overclaiming (cf. Fell & Koenig, 2016; 2020).

7.9 Overclaiming correlates- Hypotheses 16, 17 & 18

7.9.1 Method

The below analyses comprised using the following variables:

- gender (st04q01)
- socio-economic status (escs)
- school type: private *versus* public (sc01q01)
- school location size (sc03q01)

Gender was coded by a dichotomous variable with "0" indicating girls and "1" indicating boys.

Socio-demographic status of students' families was estimated using the PISA index of economic, social and cultural status (ESCS). The index is calculated on the basis of students' parents: occupational status, as indicated by the International Socio-Economic Index of Occupational Status (ISEI), highest educational level achieved (in years of schooling), family wealth, home possessions (e.g. computer, own room for participant at home, household appliances, etc.), classical cultural possessions (e.g. classical books, paintings) and educational resources (e.g. educational software, high-speed Internet

connection). The ESCS is scaled as a continuous variable with the mean of zero and standard deviation of one among the OECD countries.

School type was coded by a dichotomous variable with "0" indicating public schools and "1" private schools.

Location size was coded by an ordinal variable with five categories coding in order: village (1), small town (2), town (3), city (4), large city (5).

Pairwise correlations, multiple multi-level regression and IRT differential item functioning (DIF) analyses were used to verify Hypotheses 16, 17 and 18. DIF was estimated using the Stata -uirt- package (Kondratek, 2016/2020). In order to deepen the regression results IRT differential item functioning (DIF) analysis was conducted, with gender and escs (median split) as grouping variables. DIF is an analysis trying to detect any between-group differences in item parameters which are conditional on ability (trait) distribution in these groups. It is important to notice that DIF analysis does not serve to identifying differences in between-group trait distributions nor between-group differences in item parameters (e.g. whether item "*i"* was easier in group *f* vs. group *r)*. DIF statistics were generated from the GRM model to which the math familiarity scale was fitted in subchapter 6.2.3. Two such statistics are presented below: a) DIF significance test based on IRT likelihood ratio test (IRT-LR test), b) effect size statistic as given by IRT P-DIF effect size measure. The LR test is interpreted as any other test of this kind, however, especially in case of large samples analyses, this test tends to be liberal and indicates DIF presence even if this effect is of trivial size. Because of that DIF is always accompanied by an effect size measure which helps to decide whether the difference is of any meaningful size. The DIF effect size is most often classified to three categories: "A"- trivial difference, "B"- medium difference, "C"- substantial difference. In practice only "B" and "C" type DIFs are analysed further on (e.g. in order to revise item's contents). More on DIF analysis and formulas for the statistics mentioned above can be found in Kondratek, Skórska and Świst (2015).

Moreover, the intra-class coefficient (ICC) for overclaiming was computed in order to measure school-level differentiation of overclaiming (Jerrim et al., 2019; Merlo et al., 2018).

7.9.2 Results

The pairwise correlations analysis yielded the following results:

Variable	**math ability**	**math familiarity (reals)**	**bias (foils)**	***d'***	***c* reversed**
gender: boys	ns	-0,12	0,13	-0,10	-0,04*
socio-economic status (escs)	0,41	0,30	-0,05**	0,18	0,07
school type: private	0,10	0,07	ns	0,06**	ns
location size	0,22	0,13	-0,08	0,15	ns

*Table 39. Correlations between socio-demographic variables, math ability and OCT measures. Note: ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

According to predictions boys yielded higher OCT bias (IRT scored) and lower OCT accuracy (both IRT and SDT scored).

Socio-economic index proved to correlate positively with OCT accuracy and brought in a small negative relation with OCT bias (IRT scored) and a slight positive correlation with *c.*

Going to private *versus* public school was related with a slightly higher math ability but was in general unrelated to OCT measures, yielding only a marginal association with *d'*.

Location size was related to OCT bias- the larger the settlement in which school is located the lower the bias and the higher the accuracy (and math ability).

These zero-order correlations were further investigated in a regression equation in order to control for math ability influence on gender and escs. However, before calculating the regression models with predictors, the so-called null models were estimated in order to calculate the intra-class coefficient (ICC) for OCT measures. The ICC values produced from these models are summarised in the last row of the regression analysis table. Values for IRT indices were averaged across five PVs.

The table presented below summarises four regression equations where OCT measures were regressed on math ability and the set of socio-demographic variables commented above:

Variable	**OCT bias (foils) coefficient**	**OCT bias (*c* reversed) coefficient**	**OCT accuracy (reals) coefficient**	**OCT accuracy (*d'*) coefficient**
math ability	-0,32	0,41	0,61	0,44
complementary OCT index[119]	0,38	-0,69	0,25	-0,61
gender: boys	0,33	-0,22	-0,29	-0,24
socio-economic status (escs)	ns	0,06**	0,09	0,06**
school type: private	ns	ns	ns	ns
location size: village (referential group)	-	-	-	-
location size: small town	ns	ns	ns	ns
location size: town	ns	ns	ns	ns
location size: city	-0,17**	ns	ns	0,16*
location size: large city	ns	-0,15*	ns	ns
ICC- null model	0,02	0,02	0,13	0,06

*Table 40. OCT measures regressed (standardised regression weights) on math ability and socio-demographic variables- results of a multilevel regression with ICC values. Note: ns p> 0.05, *p<0.05, **p<0.01, no * p<0.001.*

Gender remained a significant predictor of OCT measures even when controlled for math ability and other socio-demographic variables. Boys yielded higher OCT bias, higher tendency to answer "yes" to items (as measured by *c* reversed) and lower OCT accuracy.

[119] For OCT bias (foils) it is OCT accuracy (reals), for OCT bias (*c* reversed) it is OCT accuracy (*d'*) and *vice versa.*

Higher socio-economic status was related to marginally higher OCT accuracy and also higher tendency to answer affirmatively to items (*c* reversed).

When controlled for other variables school type was not a significant predictor of OCT measures.

Location size also failed to produce a coherent set of relations with OCT measures when controlled for other variables in the model.

ICC analysis

ICC values from null models were low, especially for OCT bias measures. This result may indicate that overclaiming is mainly dependent on individual-level variables and school clustering does not have much relation to it (cf. Jerrim et al., 2019).

Differential item functioning (DIF) analysis

Among the 16 items of which the PISA 2012 math familiarity scale is composed 13 yielded a statistically significant value of the LR test (< 0.05) but only four out of them approached the non-trivial value of the DIF effect size (P-DIF > 0.25; cf. Kondratek et al., 2015). The item parameters and DIF statistics for these items is presented in the table below. All statistics for all items along with ICCs is available in Online Appendix D.

Item	LR test	p	P-DIF GR	P-DIF GF
st62q03	37,97	***	-0,238	0,243
st62q08	72,68	***	-0,292	0,305
st62q11	112,56	***	-0,385	0,397
st62q13	41,24	***	-0,303	0,314

*Table 41. Significance and effect size for DIF analysis across gender. GR= boys, GF=girls. Note: *** p < 0.001.*

Table 41. displays information about DIF significance and effect size. It is worthy to know that out of four items that have approached the "B" DIF size two are foils and two are reals. The third foil, item st62q04, yielded significant DIF but of only marginal size. Hence, four items were much easier for boys than girls conditional on the level of math ability (as measured by math familiarity scale). Thorough analysis of items' thresholds indicated that boys were more prone to claim some familiarity with the identified items (e.g. select response "2" or "3" instead of "1") but not more prone to embrace the highest responses (e.g. "4" or "5"). The information on item parameters in two groups conditional on between-group trait distribution is shown below:

Parameter	GR (boys)	GF (girls)	Item
a	1,33	1,39	st62q03
b1	-1,72	-1,29	
b2	-0,82	-0,44	
b3	0,06	0,25	
b4	0,88	1,07	
a	0,87	1,05	st62q08
b1	-0,77	-0,10	
b2	0,61	0,99	
b3	1,91	2,05	
b4	3,40	3,79	
a	0,48	0,72	*st62q11*
b1	-0,31	0,84	
b2	1,57	2,05	
b3	3,57	3,25	
b4	5,43	5,11	
a	0,69	0,67	*s62q13*
b1	-1,58	-1,04	
b2	-0,25	0,29	
b3	1,03	1,67	
b4	2,40	3,29	

Table 42. Item parameters for DIF-flagged items over gender. Note: a- discrimination parameter, b-difficulty parameter from the fitted polytomous IRT model.

Item discrimination seems roughly even between the two groups, only item st62q11 is characterised by a slightly more steep ICC among boys than among girls. It seems thus that the DIF present here is of uniform nature (difference only in item difficulties). Threshold analysis shows that these items were easier for boys than for girls conditional on trait distribution in two groups. It looks like the difference is more pronounced in the lower than higher thresholds.

Similar analyses were conducted for median-split socio-economic status (escs). It is of course a suboptimal analysis as such forced categorisation of continuous variables is not recommended, however, it is the only way to perform DIF analysis across escs.

Among 16 math familiarity scale items only two did not reach statistical significance in DIF analysis but only two approached the substantive effect size. Basic information on them is summarised in the tables below:

Item	LR test	p	P-DIF GR	P-DIF GF
st62q04	54,41	***	-0,25	0,25
st62q13	52,31	***	-0,27	0,28

*Table 43. Significance and effect size for DIF analysis across socio-economic status. GR= low escs, GF=high escs. Note: *** p < 0.001.*

Parameter	GR (low escs)	GF (high escs)	Item
a	1,27	1,04	st62q04
b1	-2,35	-1,99	
b2	-1,12	-0,87	
b3	-0,04	0,28	
b4	1,30	1,72	
a	0,57	0,52	st62q13
b1	-1,87	-0,99	
b2	-0,16	0,61	
b3	1,53	2,22	
b4	3,54	3,89	

Table 44. Item parameters for DIF-flagged items over socio-economic status (escs). Note: a- discrimination parameter, b- difficulty parameter from the fitted polytomous IRT model.

Items st62q04 ("proper number") and st62q13 ("declarative fraction") appeared again in this analysis. It seems that they were the root of most of the differences observed between low and high status groups. It is worthy to note that both items st62q04and st62q13 were embraced more eagerly by lower escs group in comparison to higher escs group conditional on trait distribution in the two groups. Comparing escs and gender DIF it is evident that there were more and more pronounced differences for gender than for escs.

As item st62q13 seemed to cause the most "trouble" in psychometric modelling of the math familiarity scale let's take a look at its ICCs, both across gender and socio-economic status group:

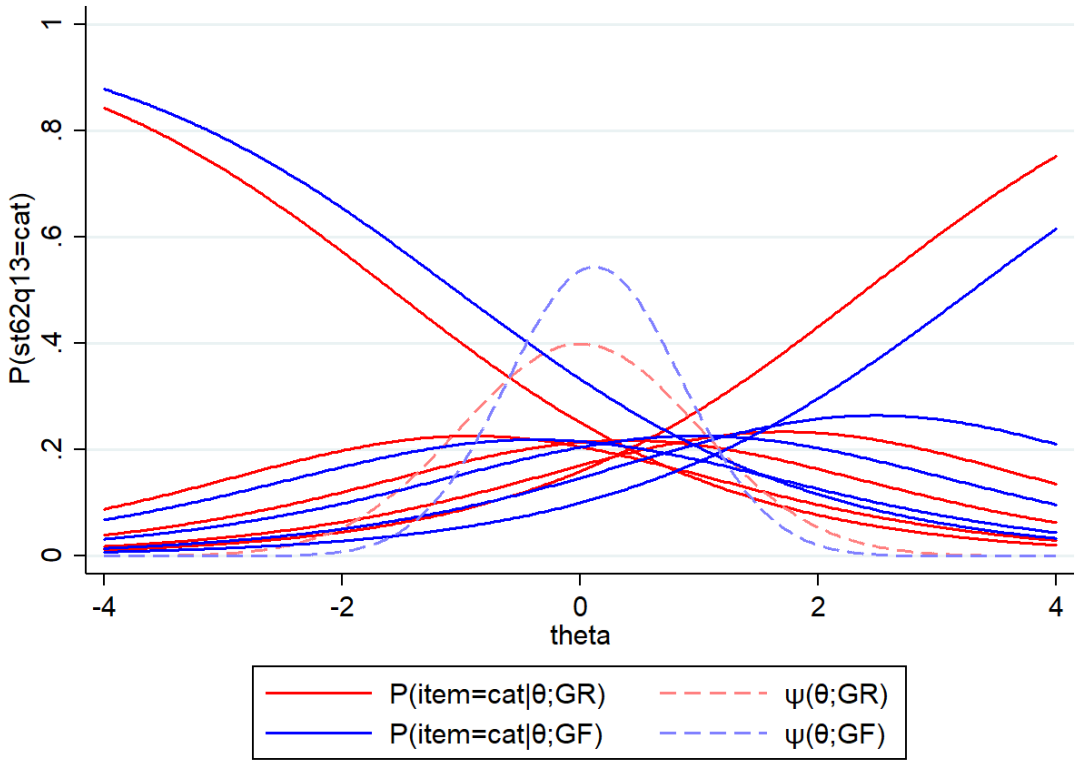

Figure 10. Item characteristic curve (ICC) of item st62q13 for DIF analysis over gender. GR= boys, GF= girls. Ψ(Θ)= trait distribution.

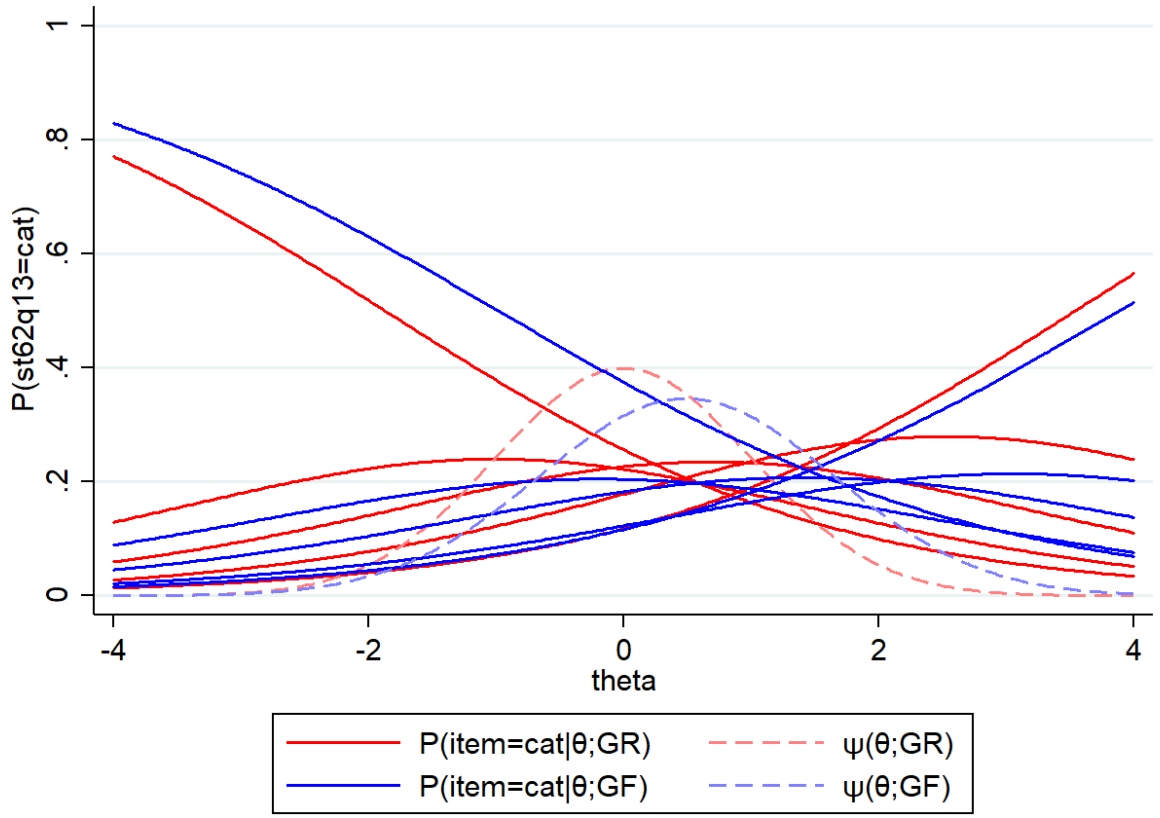

Figure 11. Item characteristic curve (ICC) of item st62q13 for DIF analysis over socio-economic status group. GR= lower status, GF= higher status. Ψ(Θ)= trait distribution.

Both figures above show two main things: a) participants responded to this item mainly using two extreme options: "1" and "5", b) boys and lower escs group, conditional on trait distribution, tended to embrace other categories than "1" more often than girls and higher escs group.

7.9.3 Discussion

Hypothesis 16 was confirmed as boys indeed overclaimed more (higher OCT bias) than girls. This effect remained significant even when controlled for math ability and other variables, hence confirming part of Hypothesis 18 as well. Such an effect is also a direct confirmation of results obtained by Jerrim et al. (2019) on a group of Anglo-Saxon participants of the PISA 2012 cycle. This result would point towards positivity bias explanation of overclaiming as there are no differences in objective math ability between the two genders, but the overall social opinion and students' attitude is that math is a "boy thing" (as it was presented in subchapter 5.3.8). This mindset may consequently lead to "knowing math stuff" being more expected from boys than from girls which could lead to overclaiming. DIF analysis showed that this difference was not very pronounced as only four items yielded significant between-group difference conditional on math ability.

Hypothesis 17 was partially confirmed as participants higher in socio-economic status did not overclaim more than their counterparts of lower status but have noted higher OCT accuracy. This result is in contrast with the Jerrim and colleagues' findings (2019) that high escs participants overclaimed more. It is hard to explain such cross-country differences[120] without generating more evidence in the

[120] Jerrim et al., 2019 used responses from the Anglo-Saxon countries participating in the PISA 2012 cycle: USA, UK (separate samples for each of the constituting countries), Australia, Canada, New Zealand, Ireland.

matter. Similarly as in the study by Jerrim et al. (2019) it was observed that item st62q04 ("proper number") displayed largest differences between groups differing in socio-economic status. Moreover, it was observed that in the Polish sample the group of higher escs yielded tendency towards more extreme responding and towards less midpoint responding. The nature of such relationship warrants further studies as up to date it is not clear where these associations stem from. It is possible that this result is linked to the results presented by Anderson and co-authors (2012) who related high social status to being (over)confident in both social relations and self-reports.

Regarding second part of Hypothesis 18: socio-economic status retained its positive relation to OCT accuracy even when controlled for math ability (and gender). These relations were, however, small in size.

Moreover, the data from the Polish sample confirmed observations by Jerrim et al. (2019) that overclaiming has low ICC, namely that it is rather evenly distributed through schools. This urges for more studies on individual-level and group-level correlates of overclaiming as the phenomenon varies greatly on individual level, seems to have almost no school-specific variance (see also subchapter 7.8) but on the other hand the cross-country (cross-culture) differences are huge (Vonkova et al., 2018), even among the countries using one language version and sharing a common cultural root (Jerrim et al., 2019).

7.10 Individual differentiation of overclaiming scores- Hypothesis 19

7.10.1 Method

Finally, an analysis of individual differentiation of overclaiming scores was planned. In order to study this topic latent class analysis (LCA) was used. LCA fits a set of latent clusters to observable response patterns in order to identify discrete groupings of participants (McCutcheon, 1987).

LCA can be used in either exploratory or confirmatory approach, here it will be applied to confirm the data pattern obtained by Yang et al. (2019; see 5.2.9 for more details on this study). This end determined the design and procedure of this study as they were tailored to resemble the procedure of Yang et al. (2019)[121]. A set of five plausible values for overclaiming, math familiarity and math ability were used as latent classes manifest indicators. Original PISA 2012 PVs for math ability (OECD, 2014b) were rescaled to the scale (0,1) (see Muthen & Muthen, 1998-2017), whereas overclaiming and math familiarity were represented by the variables created in the steps described in part 6.2.3 of this work. The data were weighted and accounted for the clustered PISA design.

LCA assumes within-class independence between the variables as a default option (Clark et al., 2013; McCutcheon, 1987). This assumption was relaxed in this analysis for PVs of the same variable as they are obviously highly correlated by design. MPlus 8.2 "mixture" analysis type was used and restricted maximum likelihood (MLR) estimator was applied to estimate the model parameters.

In order to compare differences between classes on a set of variables, class membership, based on class probabilities of the best fitting model, was saved and used as an independent variable in mixed regression analysis. Margins were used to calculate variables' means in a given latent class, controlling for the PISA design. Significance of the inter-class differences were calculated on the basis of the estimated regression coefficients.

[121] The exact similarity was probably not achieved as Yang et al.'s paper lacks certain relevant method data. The corresponding author did not respond to an e-mail with a plea to share more details.

This way of way of estimating these covariate effects is not optimal as class membership does not account for uncertainty related with class probability (Pokropek, 2016) but this way of proceeding was adopted in order to ease the calculations. Moreover, similar course of action was implemented by Yang et al. (2019) and as latent class covariates can influence posterior class probabilities it was used as a second argument not to include covariates in class identification stage. Furthermore, there were no clear conceptions on how to specify models with covariates due to lack of previous data.

7.10.2 Results

Model fit

Models differing in number of latent classes were fitted to the observed dataset. In order to decide on the number of classes to which the dataset was divided the following source of information were used: a) information criteria (AIC, BIC, sample-size-adjusted BIC), b) entropy parameter, c) Lo-Mendell-Rubin test (LMRT; Lo, Mendell & Rubin, 2001), d) theoretical model interpretation with a special regard on the solution proposed by Yang et al. (2019), e) model parsimony. The below table summarises these statistics for the estimated models:

LCA model	AIC	BIC	sample size-adjusted BIC	Entropy	LMRT *p* value	# of free parameters	log-likelihood
1 class	72538,26	72899,63	72708,99	-	-	60	-36209,1
2 class	71690,23	72147,97	71906,49	0,586	0,0001	76	-35769,1
3 class	71265,74	71819,85	71527,53	0,687	0,0008	92	-35540,9
4 class	70962,34	71612,81	71269,65	0,715	0,0786	108	-35373,2
5 class	70768,19	71515,03	71121,03	0,732	0,6693	124	-35260,1

Table 45. Model fit statistics for the LCA analysis.

The statistics gathered in the above table display situation typical for LCA- model fit indicated by information criteria increases with the increase of the number of latent classes (Clark et al., 2013; Muthen & Muthen, 1998-2017). In order to circumvent this problem the number of the classes fitted to subsequent analyses was decided on the basis of the LMRT which indicated three-class solution and theoretical interpretation which also pointed for this resolution. The four-class model consisted of very similar classes as the more parsimonious model with the additional, fourth class being simply split of one of the classes identified in the three-class solution. Moreover, these additional, offspring classes had very similar profiles suggesting that the differentiation between them was largely spurious (Clark et al., 2013). Finally, and most importantly, the three-class solution directly corresponded with the previous model adopted by Yang et al. (2019). Hence, three-class solution was elected and subjected to further scrutiny[122][123].

[122] The value of entropy parameter for this model, 0.687, also indicated that this model assured sufficient class separation. It is worthy to note that this value is higher than the entropy value in the accepted model in the study by Yang and colleagues (2019) where it amounted only to 0.549.

[123] Complete results for the four-class solution are available in Appendix F.

Latent class profiles

The below table summarises basic characteristics of the three classes emerged:

Class	Estimated posterior probability	N based on most likely membership	Proportion based on most likely membership	Average latent class probability	Average latent class probability (based on most likely membership)
1	0,42	1271	0,42	0,865	0,857
2	0,09	259	0,08	0,856	0,787
3	0,49	1521	0,50	0,843	0,862

Table 46. Estimated posterior probabilities for class membership.

Two largest classes accounted for more than 90% of participants. The comparison between estimated probabilities based on posterior distribution probability and most likely class membership proved to be similar to each other which positively attests for the stability and adequacy of the chosen solution.

Means of the indicators in given classes are presented in the figure below:

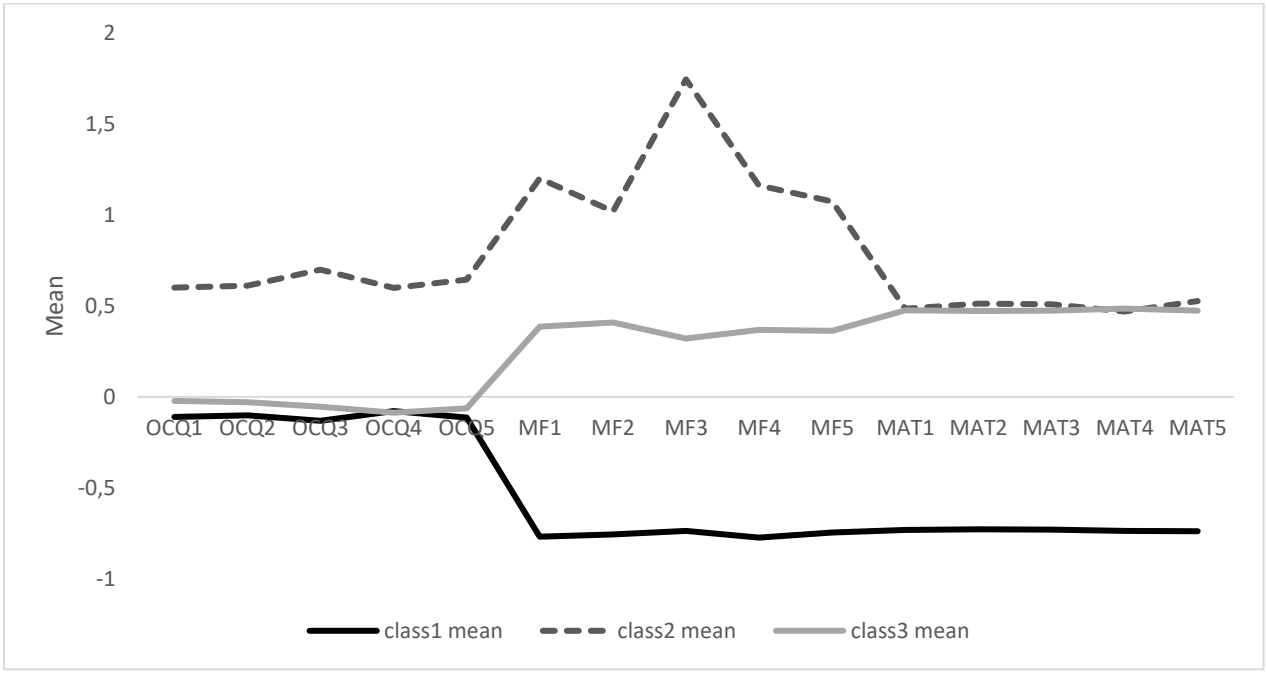

Figure 12. Latent classes' profiles

The smallest of all three classes class 2 (representing only 9% of participants) seems to resemble the "overclaiming" class discerned by Yang et al. (2019). This group is characterised by the highest overclaiming and math familiarity scores but its scores on the objective math ability test do not seem to corroborate such a high opinion on their abilities. Class 1 can be described as low achieving class with an average score on overclaiming. This pattern of results seems to warrant the hypothesis that

some of the high overclaiming scores were noted by participants endorsing midpoint response category when responding to foils in a way to, allegedly, show their confusion with the "odd items". This hypothesis seems corroborated by low scores of this group on math self-report and math ability alike, indicating their overall accurate self-description. Finally, the largest class in the model, class 3, yielded response pattern that could be commented as class of reasonably high mathematical ability, low overclaiming and adequate claims on the math familiarity scale (accurate self-report regarding math abilities).

Inter-class differences on relevant variables

Mixed regression models were used to study inter-class differences on a set of predefined variables. Variables were qualified on the basis of two premises: a) Yang et al. (2019) paper, b) correlations with OCT measures obtained in the previous analyses (see subchapters 7.2-7.9). The below table summarises findings provided by these models and presents marginal means in each class:

Variable/class	1 (low ability, accurate self-report)	diff	2 (overclaiming class)	diff	3 (high ability, accurate self-report)	diff
Male	0,52	3	0,49	-	0,44	1
socio-economic status	-0,47	2,3	0	1	0	1
self-efficacy	-0,61	2,3	0,64	1,3	0,41	1,2
internal failure attribution	-0,2	2,3	0,26	1,3	0,08	1,2
work ethic	-0,29	2,3	0,31	1	0,16	1
learning behaviour	-0,2	2,3	0,39	1,3	0,11	1,2
math interest	-0,26	2,3	0,48	1,3	0,23	1,2
instrumental motivation	-0,22	2,3	0,4	1	0,33	1
math anxiety	0,49	2,3	-0,69	1,3	-0,36	1,2
self-concept	-0,54	2,3	0,64	1,3	0,27	1,2
attitudes activities	-0,13	2,3	0,22	1	0,08	1
attitudes outcomes	0	2	0,17	1,3	0	2
school belonging	0	2	0,26	1,3	0	2
teacher-student relations	0	2	0,17	1,3	0	2

Table 47. Marginal means for latent classes' covariates. Note: all variables on a standardised scale with SD=1; diff- yields number of the class or classes that differ significantly from the class indicated in the column header to the left. All variables, save gender, were measured on a scale with mean equaled zero and standard deviation of one. Gender is given as a percent of boys in a given class.

The “low skilled, accurate self-report” class comprised larger percent of male students than any other class and was characterised by lower than average scores on most of the variables, including socio-economic status. This group noted average results only on scales measuring school attitude (attitude towards outcomes, school belonging and teacher-student relations) and a markedly above-average score on math anxiety.

The ”overclaiming” class did not differ in case of gender and socio-economic status from the “high skill, accurate self-report” class. It did noted however significantly higher scores on most of the self-reports, yielding a very desirable profile of a highly competent and socially successful student.

The last class yielded a favourable profile but to a much lesser extent than the “overclaiming” class. It is noteworthy that the last class did not differ in the scores describing social relations from the “low skilled” class but noted lower scores than the “overclaiming” group.

7.10.3 Discussion

Comparing the results obtained to the findings of Yang and colleagues (2019) it has to be said that in general the pattern of findings is similar. The “disengaged” class from the Yang et al. study largely resembles the “low, accurate” class from this study, with the exception that this class grouped around 20% of participants in the Yang’s study in comparison to 42% in this study. Moreover, the “overclaiming” class in the former paper counted more than 20% of the students which was more than twice the number obtained in the latter work (only 9%). Finally, the “high, accurate” class was also marginally larger in the Yang’s study (around 60%) than in the above presented work (almost 50%). This shows that the main difference resides in the “overclaiming” class being much larger and the “low, accurate” *alias* “disengaged” class much smaller in the American in comparison to the Polish sample.

Regarding the variables regressed on class membership the inter-class differences seem similar when comparing the two studies. It seems that the LCA confirms that males are characterised by higher overclaiming and that overclaiming is a feature related rather to higher, not lower, social status (cf. Jerrim et al., 2019). The latter result reminds the findings obtained by Anderson et al. (2012) who pointed to the role of self-assurance, even if not congruent to reality, in winning social position. In general the “overclaiming” class yielded similar profiles in American and Polish samples, with a somewhat more extreme, desirable profile of the “overclaimers” in the latter one. It is probable that the Polish “overclaiming” class, being smaller than the American one, comprised only “die-hard overclaimers”.

Some differences between the studies can be most certainly attributed to the cultural differences between the USA and Poland. Large proportion of the American students in the “overclaiming” class found by Yang and co-workers (2019) is not surprising as Jerrim and colleagues (2019) assessed that students in the USA overclaimed their mathematical skills to the greatest degree among all the Anglo-Saxon countries in PISA[124]. However, Vonkova et al. (2018) obtained similar values of accuracy and exaggeration indices for the two countries compared here. Thus, it seems that data at hand enables to predict that both countries should yield a notable “overclaiming” class but do not offer much suggestion on how to explain any observed differences between the USA and Poland[125].

Interestingly, the “low, accurate” class yielded similar OCT scores but lower math familiarity scores than the “high, accurate” class (this result is practically identical when compare Yang’s study and this work). This effect can be interpreted just as the floor effect but more scrutiny would be warranted to

[124] Save Canada which noted the same results as the USA.

[125] It is also possible that at least some of the differences are attributable to the slight methodological discrepancies between the two studies instead of substantial factors.

exclude other explanations, e.g. MRS. Studies with the use of other data sources, e.g. qualitative interviews or eye-tracker data, would bring great help to understand the differences between the classes more fully.

The performed LCA yielded three classes, largely similar in profile to the classes obtained by Yang and colleagues (2019) on a different sample. The "overclaiming" class was characterised by unwarranted high level of self-report math skills and math-related abilities, yielding a very desirable and (overly) positive profile. However, this group did not have higher math abilities in comparison to the other class which noted high abilities but yielded much more moderate self-report scores. Nevertheless, the "overclaiming" group counted only 9% of the participants, a proportion much lower than often cited 30% in the literature (or 20% in the Yang's study; see subchapter 4.4 for a detailed discussion on this topic). It is probable that more studies with the use of non-US samples will help to shed more light on this proportion. It is notable that one of the very few studies that compared accuracy of self-report in comparison to objective test on a non-US sample also obtained a slightly lower proportion of overclaimers (16%; Koniewski et al., 2019). More research is warranted on cross-cultural differences and covariates of overclaiming. LCA studies are also recommended as they seem to bring interesting insight into the investigated matter. The above studies development, e.g. with the use of factor (IRT) mixture models, covariates or multilevel LCA (Henry & Muthen, 2010) is also warranted.

Chapter 8- SUMMARY AND CONCLUSION

8.1 Results' summary

Suppression effect and OCT scoring

Obviously, the analyses performed could not answer all the questions related to the overclaiming technique. However, the approach of small research steps (Kalton & Schuman, 1982) enabled to inform on certain research problems, advancing a middle-level theory and laying ground before any further research attempts in this area.

First of all, it was determined that including OCT bias index in a regression equation indeed results in a suppression effect, namely in enhancement of predictive validity of a subjective math ability on an objective math test. However, in contrary to predictions, the revealed pattern was not of classical suppression but of mutual suppression in which OCT bias was negatively related to math ability in a zero-order relation and where both OCT bias and OCT accuracy indices boosted their predictive validity when introduced into regression equation.

Comparing to the radical improvement of self-reports predictive validity and interpretability achieved in cross-country context (Khorramdel et al., 2017; Kyllonen & Bertling, 2013) the results of suppression analysis for the within-country, inter-individual context was much more modest.

Nevertheless, the enhancement of regression weight was statistically significant but increment in its size was rather moderate, especially in case of IRT indices where the regression weight risen only of about 7% and R^2 increased by 10%. These values were dwarfed by the increment noted by SDT indices, where predictive weight of the subjective math ability was boosted by 200% and R^2 was almost doubled. Common sense indices brought increments of a medium size, somewhat between IRT indices and SDT measures.

Important conclusions were reached regarding OCT scoring. Three scoring approaches were compared: IRT indices, SDT indices and "common sense" indices (Paulhus et al., 2003; Vonkova et al., 2018). It was determined that indices differ in their exact interpretation and that misinterpretation of the *c* parameter is very likely due to the common confound caused by the minus sign in the equation. This minus sing was commonly dropped in some studies but a browse through the literature indicates that not in every study it was reported properly which often could distort results' interpretation.

Moreover, the common sense indices do not seem to offer equal measurement characteristics as SDT indices. More comparative research in this topic is warranted, but for now the common sense indices should be used with extra care. Furthermore, it is not known, as it was not tested in this work, how a wide family of *ad hoc* OCT indices, often used in the field, would behave in such analyses. Basing on psychometric knowledge of the properties of sum scores (von Davier, 2010) or difference indices (Zumbo, 1999) their use should be limited in favour of more advanced and theory-based scores such as IRT or SDT indices. What is more, it was revealed that including interaction between SDT statistics *d'* and *c* leads to further increase in predictors' criterion-related validity and variance accounted for by the model (as measured by R^2). This issue also calls for future exploration as interactions between OCT bias and accuracy scores were not analysed in previous studies. Bypassing this term in regression equations could result in model miss-specification (specification error) and biased estimates. Furthermore, as it was evidenced in the subchapter 7.2, using different OCT scoring may lead to different research conclusions.

Interestingly, it seems that SDT indices' interpretation as pure measures of OCT bias I and OCT accuracy (*d'*) needs to be revised. Both indices most likely confound information on accuracy and bias. Especially the *c* statistic should be interpreted more as warranted by the SDT theory- as a tendency towards certain type of responding, which, in case of responding to Likert-type rating scales, not necessarily equals to bias. It appears that SDT measures should be only interpreted together, as they do not make much sense apart from each other. Moreover, as it can be assumed from the interaction analysis shown in the subchapter 7.1, that any responses yielded from respondents should be interpreted in a 2 x 2 matrix for a more precise interpretation. It is suggested that truly accurate respondents are those scoring high on both *d'* and *c* (at least in the context of rating-scales), whereas participants scoring low on *d'* but high on *c* are those yielding positive bias. This issue is related with interpreting correlation between OCT bias and accuracy. These value differs greatly across indices (see subchapter 7.5) but also across studies and samples (e.g. Ziegler et al., 2013).

Another, as yet unaddressed issue with SDT indices, is their reliance on assumptions derived directly from the Signal Detection Theory, e.g. normal distribution of both signal and noise and equal variances characterising these information. It is unlikely that breaking these assumptions (which is evident in case of overclaiming tasks applied to rating scales) is without the consequences for psychometric quality and interpretability of SDT measures in the OCT field. Such a direct transplant from a very different theory is a welcome contribution, broadening the field of social sciences, however, it is advised to deepen our knowledge on consequences of using these indices for rating scales. It is likely that in this context most of the SDT's basic assumptions are broken (cf. Goecke et al., 2020). Moreover, SDT's use in the context of social sciences was lately criticised as the theory was deemed an "oversimplified metacognition model" (Paulewicz & Blaut, 2020). New extensions of SDT for non-dichotomous data are needed and may prove an important addition for response biases research. Such extensions for Bayesian models (Paulewicz & Blaut, 2020) and multilevel techniques (Wright, Horry & Skagerberg, 2009) were proposed lately. Both methods await their validity studies in the framework of social sciences, in particular in the field of self-reports.

Mechanisms of overclaiming

Dunlop and colleagues (2017) proposed four possible mechanisms causing overclaiming: a) motivated positivity bias resulting in self-enhanced self-presentation, b) faking/impression management/lying, c) memory bias, d) careless responding. This proposition was taken as a guideline to plan hypothesis testing that could provide indirect evidence on the veridity of these mechanisms. As no measure of deliberate response distortion (e.g. faking) was available in the PISA dataset this option remained untested. However, it was well evidenced that such an explanation of overclaiming is neither necessary (Greenwald, 1985; Kozielecki, 1981) nor likely to explain overclaiming in low-stakes assessments (Paulhus, 2002; Ziegler, 2015).

Memory bias account was partially tested by analysing relation between math ability and OCT indices. This analysis enabled to compare the overgeneralisation account, claiming that participants higher in ability should overclaim more due to interference from other terms stored in the long-term memory (Kuncel et al., 2012), with the metacognitive account, linking overclaiming with insufficient control over one's knowledge, which predicted that participants lower in ability should yield larger bias (Atir et al., in preparation; Paulhus & Dubois, 2014). Moreover, relations between openness and OCT bias were tested, in order to inform the discussion on memory bias further on, as positive relation between openness and bias would offer support for the overgeneralisation account, whereas no relation or a negative one would disconfirm it.

The results presented in the subchapter 7.2 negated the overgeneralisation account as math ability was related to less overclaiming, instead of more, thus offering support for the metacognitive account. Furthermore, openness was related only to OCT accuracy (positively) but failed to produce any significant relations to OCT bias as measured by IRT score for foils. It is important to note, that in this analysis *c* and IRT score for foils again were linked with discrepant results- openness or self-efficacy were not related to the IRT score but were positively related to *c*. Similar result was obtained by Bertsch and Pesta (2009) where IQ did not correlate with *c*, but was negatively associated with mean score for foils.

These results seem to corroborate studies finding negative relation between cognitive ability and OCT bias (e.g. Atir et al., in preparation; Franzen & Mader, 2019; Kyllonen & Bertling, 2013; Paulhus & Dubois, 2014), against opposite propositions (Kuncel et al., 2012) or null results (Bing et al., 2011). Moreover, no relation between openness and OCT bias was found, further disconfirming the overgeneralisation hypothesis. The obtained results showed positive relation between openness and OCT accuracy but no relation with bias, yielding similar pattern to many previous studies (e.g. Dunlop et al., 2017; Ludeke & Makransky, 2016; Swami et al., 2011). However, positive relation between openness and OCT bias was found by some researchers, e.g. Ziegler and co-authors (2013). It is possible that these relations differ across overclaiming tasks, e.g. domains they measure, and specific facets of openness (Christensen et al., 2019; Schwaba et al., 2019). Future studies should further explore this topic as the relations between personality and cognitive functioning may possibly help to clarify relations between response process in OCT and a general style of forming judgments (both self-judgements and judgements about OCT terms). Atir and colleagues (in preparation) claim that using automatic, fast decisions when assessing one's familiarity with OCT terms is positively related with OCT accuracy and negatively with bias, whereas deliberative processing yields more bias. Moreover, these researchers also claim that participants noting low bias scores do not generate many associations with foils. These results further disconfirm overgeneralisation theory and suggest that the relation between openness (more associations with foils) and OCT bias predicted by some accounts (Ziegler et al., 2013) seems at best characteristic of some very limited contexts only. Furthermore, it would be interesting to deepen our understanding of how participants form judgements about themselves and about OCT items. Integrating Kozielecki's (1981) and Lewicka's (1978) research on judgements formation with the recent research presented by Atir et al. (2015; in preparation) and Klimoski and Hu (2011) seems especially promising.

Finding evidence that overgeneralisation account does not hold in front of the above results of course does not eliminate the memory bias hypothesis entirely. First of all, the supported metacognitive theory also contain memory-related component (although to a lesser degree than the overgeneralisation account), moreover, other non-motivated memory biases may play role in overclaiming, e.g. hindsight bias. Muller (2019, Study 3; see also Muller & Moshagen, 2018) claims that OCT bias is due to reconstruction, not recollection bias, thus participants embrace foils, because they are genuinely convinced that they have previously heard about them[126]. It is important to note, that this process should be even more elevated in the PISA 2012 OCT version, due to the foil construction rules- two real math terms were joint to form a non-existent term. However, the role of foils' verisimilitude is scarcely known- Franzen and Mader (2019) reported that more plausible foils were claimed more often than less plausible ones but Calsyn and others (2001) showed that using "more familiar" foils led to higher claiming of reals, but NOT foils. Atir and co-workers (in preparation) presented results were foils embedded in one list with more known (easier to claim) reals were embraced more often than the same foils surrounded in a list by less known (more difficult to claim)

[126] Cf. with the results presented in subchapter 7.7.

reals. The researchers attributed this result to the assimilation-contrast effect (Lord, Ross & Lepper, 1979). Other possible explanations of such an effect entail engagement of cognitive control in more difficult list of items, similarly, as it is done in cognitive control tasks (e.g. Stroop or flanker tasks) or semantic priming (Tillman & Wiens, 2011). These accounts were impossible to comment on in this work due to lack of data in the PISA dataset, however, they call for experimental verification in the nearest future.

In the subchapter 7.2 it was also evidenced that subjectively-measured math ability produced distinct relations with OCT measures in comparison to objectively-measured math test. This result points to further scrutiny over positive relations between OCT bias and self-reported math-related traits, evidenced in previous research (e.g. Atir et al., 2015; Jerrim et al., 2019). It is possible that such relations have stylistic rather than substantive explanation as *c* should not be treated as a pure OCT bias index. Moreover, such relations seem to vanish when controlled for *c*, which further suggests that they are based only on spurious variance (e.g. method variance, RS, etc.). However, it is up to further studies to determine whether such self-concept scales could possibly act as a moderator of the OCT – math ability relation.

Another analysis concerned the alleged domain-specificity *versus* domain-generality of overclaiming. The results gathered point rather to a domain-general character, as OCT bias and accuracy measures were related to all three PISA domains (math, reading, science). It is thus assumed that it is rather general cognitive ability than specific domain knowledge that is related to OCT scores (cf. Atir et al., 2015). This results is in some contrast with the fact that OCT items are notorious for low inter-domain correlations, namely foils and reals from different domains, e.g. math and fashion, tend to correlate at best modestly with each other (e.g. Calsyn et al., 2001; Franzen & Mader, 2019). This is a somewhat puzzling result that could not be addressed in this work due to only one domain (math concepts) of which the PISA 2012 OCT version consisted.

Another set of hypotheses was related to OCT bias as a result of motivated positivity bias, e.g. self-enhancement. To this end scales measuring individual math-related attitudes (importance, motivation, interest, anxiety) were correlated with OCT measures. The results obtained may be carefully interpreted as supporting the role of positivity bias in driving OCT scores. Math-related attitudes correlated positively with bias (both IRT- and *c*-scored) and were not related to OCT accuracy as measured by *d'*- a pattern indicating relation to response bias (Paulhus et al., 2003). Moreover, external attribution to failure correlated positively with OCT bias, just as it was found in the works of Paulhus and John (1998) who related external locus of control to narcissism and motivated self-enhancement. Another corroboration for positivity bias being source of OCT scores came from the positive correlation of the school outcomes scale[127] with OCT bias (IRT-measured) and negative with OCT accuracy (*d'*). This relation could be explained by a socially desirable responding tendency, as embracing such items is obviously against the social consensus that school is a wonderful place for each and every young person. Hence, there is a certain evidence that OCT scores are indeed at least partially correlated with positivity bias. The effects are small in size though, which warrants further research with the use of locus of control, failure attribution and school outcomes scales in order to replicate the above findings.

Motivated biases embrace not only exaggerating positive traits but also concealing vices, sometimes this effect is called defensiveness (Paulhus & John, 1998). Analysis of school-related rule violations, as perceived and reported by students and principals, was performed in order to test relations of such processes with OCT measures. Moreover, discrepancies between principals' and students' views on

[127] Which measured attitudes to items like "School has been a waste of time."

school discipline were analysed to a similar goal. The results brought in another slight evidence for a relation between socially desirable responding and OCT bias as the respondents seem to overclaim not only their familiarity with math concepts but also their school belonging and relations with teachers. Both concepts have obvious SDR, even sensitive questions connotations, thus a positive association with OCT bias (and negative with *d'* measure of OCT accuracy) falls exactly into the established pattern of positivity bias in OCT scores (negative or null relation with accuracy, positive with bias) and, most importantly, into a theory-predicted model of relations.

However, no principal-reported data correlated with any of the OCT scores, neither did difference between school headmasters' and students' view on school discipline. Such relations are worthy of future studies, e.g. with the use of school administrative data, e.g. on school-level of truancy, discipline or safety. Use of parent- or teacher-reports also seems a promising idea to explore in future research projects, e.g. with the use of MTMM matrices. Moreover, it is also up to future endeavours to disentangle certain contradictions between the school-level discipline breaking (no relation to OCT bias) with country-level rule violation (more violation, more OCT bias; cf. Fell & Koenig, 2016; 2020; Fell et al., 2019). There is also a certain caveat to overly-optimistic interpretations of the relations found- as they are small in size, their stylistic explanation is still viable (cf. Khorramdel & von Davier, 2014). Same comment appertain to the analysis of school-level (contextual) effects of social norms and school-level rule breaking which both failed to yield coherent relations with OCT measures.

School accountability procedures turned out to be unrelated to OCT scores. The yielded correlation with math ability (both self-reported and objectively measured) is a very interesting effect but its explanation falls beyond the scope of this work.

Another mechanism possibly standing behind OCT scores that was suggested by Dunlop and others (2017) concerned stylistic responses (RS) and careless responding (C/IER). C/IER indices were found to be related to OCT scores and it is possible that certain participants, those indicated as yielding aberrant, outlying response vectors, may indeed generate their OCT scores unknowingly, simply as a by-product of careless responses. It was also evidenced that C/IER indices moderate OCT scores characteristics, e.g. size of the correlation between bias and accuracy. Person-fit measures seemed to interact with OCT scores to the highest degree, thus it seems that further research should especially concentrate on these measures and their relations to OCT (see similar results in Ludeke & Makransky, 2016). However, the framework of careless responding is still rather in its infancy as many methods are under-researched (cf. Fronczyk, 2014) and there are more unknowns than knowns in the field, despite tremendous advancements done recently (e.g. Meade & Craig, 2012).

PISA rotation design enabled to conduct a "natural experiment" regarding significance of respondents' fatigue on OCT measures. However, the fatigue condition was not related to OCT bias but yielded a significant interaction with OCT accuracy- fatigue decreased the OCT accuracy's predictive validity on math ability. Fatigue was also related to C/IER indices, although the size of the zero-order correlations produced was small. Further empirical research in this very much forgotten are of survey methodology is needed (Herzog & Bachman, 1981; but see recent revival of this topic, e.g. Yan, Fricker & Tsai, 2020).

IRTree approach towards measuring response styles (Böckenholt, 2012; 2014; Khorramdel & von Davier, 2014) was adopted in this work. The estimated values for ERS and MRS produced significant relations with OCT measures, however the pattern was different as predicted. ERS was related positively to OCT accuracy whereas MRS was related positively to OCT bias (more MRS, more bias). These results are not in concert with the findings of Dunlop et al. (2019) where ERS was related positively to both OCT accuracy and bias, a result predicted also in this work. However, different RS generation as well as dissimilar OCT scoring precludes any decisive comparisons. It seems though, that

specific ERS coding elected by Dunlop and others (2019), where only the highest positive category was coded as extreme (cf. Khorramdel & von Davier, 2014), was the main motor of such pattern of results. Different RS mapping matrices could be used in order to test their relation towards OCT scores (Böckenholt, 2014; Falk & Cai, 2016; Plieninger, 2020a). Counter-predictive results may also stem from the specificity of the OCT task used in PISA 2012. Participants may yield stylistic responses to foils as an expression of confusion towards terms they never heard of. This result calls for further analyses of these relations, both with the use of RS indices but also with more qualitative approach, e.g. cognitive interviews that could help to solve the mystery of how participants react to foils and what meanings are ascribed to responses to these very specific survey items.

Altogether, the possible origin of OCT scores is most probably heterogenic, as memory biases, metacognitive processes, positivity bias (both self-enhancement and SDR) and stylistic variance (C/IER, MRS) all play important role as OCT scores' likely originations. Their disentanglement and quantification (what is the dominant source of OCT variance?) is a task for future studies. However, the analyses presented in the subchapter 7.7 indicate that before such endeavours will be undertaken it is warranted to research consequences of certain OCT design characteristics such as number of items, number of domains, foils-to-reals ratio, foils construction rules and foils risk of confusion (Franzen & Mader, 2019). More knowledge acquired about this initial step would aid to obtain more interpretable, stable and psychometrically sound results in future studies (see also Steger et al., 2020).

Establishing overclaiming's nomological network

Male participants were found to overclaim more and obtained lower accuracy in the PISA 2012 OCT than female respondents. This pattern of relations seems prevalent in the literature, as it was already established by Noelle-Neumann (1974; men yielded less conformist responses), Bishop and others (1986; men offered more opinions on non-existent topics) or Ones and Viswesvaran (1998; men achieved higher distortion of scores in faking paradigms). Jerrim and colleagues (2019) also confirmed this relation on the PISA 2012 data for the sample from Anglo-Saxon countries.

However, the precise nature of this relation remains elusive. Nevertheless, in contrast to e.g. C/IER and RS relations with OCT, associations between gender and OCT are much better researched. It seems that there are three main group of effects that are possible explanations to more overclaiming among men: a) domain, b) personality and c) memory effects.

In case of the first group of effects it is warranted to conclude that importance, desirability and relation to identity of a given domain all predict more overclaiming. In example, Rynko and Palczyńska (2018) obtained higher overclaiming among men in comparison to women in math-related tasks, but no such effects in ICT-related tasks. Moreover, math, as a part of academic world, is one of the typical examples of an agentic domain, which are overclaimed more by men in comparison to higher bias in communal domains among women (Paulhus & John, 1998). Furthermore, math is regarded as a "boy thing" among Polish students (cf. Baczko-Dombi, 2017; Cipora et al., 2015; 2018), hence, it is a domain more central to male self-identity which predicts higher OCT bias in this group, precisely as it was shown in this work. This line of inter-gender research is also able to inform on the domain-specific (Ziegler, 2011) *versus* domain-general (Paulhus, 2002) character of overclaiming. It is justified to assume that overclaiming, as a function of positivity bias, may take place in any domain but domain characteristics are potential moderator of its appearance and size.

Another possible solution of the gender differences in OCT remains in the personality framework. Paulhus (1984) identified personality basis of overclaiming as a "convolution of high self-esteem and low anxiety", a profile characterising men more commonly than women. Moreover, higher levels of overclaiming were linked with high, but unstable self-esteem (Kernis, 2003), competitive worldviews

(Paulhus & Trapnell, 2008; Schilling et al., 2020), risk-taking (Ziegler et al., 2013), grandiose narcissism (Zajenkowski et al., 2019) and intellectual humility, which Krumrei-Mancuso and others (2019) labelled as "healthy confidence of one's knowledge". Most of the inter-gender differences in these traits point into male respondents as more prone to overclaiming.

Finally, gender differences in memory effects also seem to indicate towards higher proneness of men to memory biases (which may lead to overclaiming, see Muller, 2019). First off, women are characterised by better memory of details, e.g. prices (Barzykowski, Leśniak & Niedźwieńska, 2010). Moreover, women's memories tend to be more vivid, detailed, emotional and more focused on interpersonal relations than men's (Niedźwieńska, 2003). Women are also less susceptible to memory biases in test probes recollection (Bridge, 2006) and are better at recall of lists of objects presented (Baer, Hayes, Trumpeter & Weathington, 2006). All these differences may eventuate in higher OCT bias scores obtained by men.

Another variable related to OCT bias was the socio-economic status of participants' families which was related positively to OCT accuracy (small, but robust effect) and was not related to IRT score of OCT bias. These results are not in parallel of what Jerrim et al. (2019) found in the Anglo-Saxon sample where socio-economic status correlated positively with OCT bias. More research is needed to solve these differences and inform on the role of social status and response biases. One theoretical link predicts positive relation between higher social status and overconfidence which may be related to OCT bias as well (Anderson et al., 2012). Moreover, school location size (village, town, city) and school type (private *versus* public) were not related to OCT measures.

Distribution of variance between levels

OCT measures yielded considerable variance on the individual level, however, as was evidenced by the analyses presented in subchapters 7.7-7.9, they yielded almost no variance at the school-level. Moreover, this effect confirms findings of Jerrim et al. (2019) for different subsample of the PISA 2012 dataset. Thus, it seems that overclaiming is almost completely independent from the school-level variance and that overclaimers are not clustered together in the same schools. More research is needed in this field, especially as the between-countries differences in overclaiming seem huge (Vonkova et al., 2018), moderating the predictive validity of math self-reports on the country-level (Kyllonen & Bertling, 2013) and depending on country-level variables as e.g. GDP, corruption level and rule-breaking tolerance (Fell & Koenig, 2016; 2020; Fell et al., 2019). It certainly warrants further investigation why no school-level norms seem to be related with OCT bias. It is possible that simply school is not a medium of OCT-related variance which may be entirely independent from educational part of lives of (junior) high-school students or may group on different levels of analysis, e.g. classrooms. Research on bullying and school discipline shows that indeed classroom level of analysis may be more suitable to gauge any school-related social relations (Dijkstra, Lindenberg, & Veenstra, 2008; Pozzoli, Gini & Vieno, 2012). Additionally, twin studies on both RS (Kam et al., 2013; Melchers et al., 2018) and OCT (Luo et al., 2019) proved that only few percent of these measures' variance was related to the shared environment. Perhaps this is another suggestion that school, which is a good example of shared environment for many twins, is not a medium of response biases variance.

Internal structure and individual differentiation

The internal structure study yielded the bifactor structure as the best fitted to the data. General factor was interpreted as math ability while two specific factors were tentatively given a technical interpretation, namely, they were interpreted as method instead of substantial factors. Such interpretation in case of bifactor models is common and leads to accounting for method variance, which is one of the assets of the bifactor model (Reise et al., 2016). Examples of such method, technical

factors comprise positive vs. negative wording of items (Marsh, Scalas & Nagengast, 2010; Woods, 2006) or non-substantial item content differences (Brouwer, Meijer & Zevalkink, 2013).

During the analysis three model solutions were assumed: a) items grouped into reals and foils specific factors, b) items grouped into dyadic and single factors (some items were compounds consisted of two words, some of just one), c) items grouped into hard and easy factors. The last option yielded the best model fit of the three solutions (and an acceptable model fit overall). Regarding the three options tested, the first had an obvious substantial interpretation of specific factors (different response processes used to answer reals and foils), the second had only method interpretation (e.g. linguistic processing could be different in case of dyadic and single items), whereas the third option can be given either method or substantial interpretation. Item difficulty is a known moderator of self-report validity, with hard items being more overclaimed than easy items. Altogether, the general factor accounted for almost 70% of the total common variance, whereas the specific factors accounted for slightly more than 20% of it. The remaining variance can be attributed to random (or unmodelled) sources. Nevertheless, this analysis showed that OCT cannot be modelled by a unidimensional solution.

It is interesting that some foils were responded to like they were just difficult reals (item st62q04), while there were also some reals that were responded to like if they were foils (e.g. item st62q08). Similar situation was observed by Steger et al. (2020) where respondents overclaimed they knowledge to real but very difficult items. Many methodological differences between the two studies preclude from drawing any firm conclusions from the Steger et al. (2020) study but it seems warranted to check if any difference can be obtained in an OCT that would include foils, reals and very difficult reals. Such an OCT version could also contain items without correct answer, items with an explicit “don’t know” response option (category) or an additional answer confidence measure (cf. Anderson et al., 2012; Bensch et al., 2017). This study design could prove to be an interesting development of the standard OCT version with lesser risk of memory biases and RS or C/IER variance. Moreover, it would be very interesting to verify who will be ready to admit his/her ignorance by choosing the “don’t know” option?

LCA study discerned three latent classes, one of which can be interpreted as an “overclaiming” class. This grouping was characterised by high foils embracing and overly positive self-report that was not warranted by their objective abilities. Moreover, also in other self-reports this group yielded overly positive responses, unlikely given their objectively-measured characteristics. This analysis confirmed relation between overclaiming and gender as well as socio-economic status. However, it also brought some discrepant results regarding relation of math ability and OCT scores. When the whole sample is analysed together math ability is evidently related negatively to OCT bias, however, the “overclaiming” group is characterised by higher, not lower than average math abilities. Nevertheless, when this relation was modelled inside of every latent class the relation still held, indicating that within the “overclaiming” class more overclaiming is still related with lower math abilities.

The hypotheses and the results of their verification are briefly summed up in the table below:

Hypothesis	Verification
1. OCT is a suppressor of the self-report-objective measure relation.	*Confirmed:* OCT bias measure indeed acts as a suppressor, enhancing the relation between self-report and objective measurement of skills
2. Objective domain ability is related to overclaiming.	*Confirmed:* Objective ability is related positively to OCT accuracy and negatively to OCT bias.
3. & 4. Subjective domain ability is related to overclaiming.	*Failed:* Subjective ability assessments were not related to OCT bias.
5. & 6. Domain desirability is related to overclaiming.	*Confirmed:* Higher domain desirability (e.g. higher motivation to learn maths) was related positively to OCT bias pointing to a motivational character of OCT bias.
7. Locus of control is related to overclaiming.	*Confirmed:* External locus of control was positively related to OCT bias, but the magnitude of relation was very small.
8. & 9. Withholding negative information is related to overclaiming	*Failed:* Rule violation measures were not related to OCT scores.
10. School pressure on domain achievement and overclaiming.	*Failed:* School-related variables were found to be mostly unrelated to OCT scores which presented only a very limited variance across schools.
11. Careless responding and respondents' fatigue are related to overclaiming.	*Confirmed:* Careless responding was moderately related to OCT scores: negatively to accuracy and positively to bias.
12. Response styles are related to overclaiming.	*Partially confirmed:* Relation between response styles and OCT scores was found to be very dependent on technical aspects of such analysis. More research is needed in order to generalise these findings.
13. Latent structure of overclaiming scale: two factors will emerge, one for reals, one for foils.	*Partially confirmed:* Bifactor solution with two specific factors (hard vs. easy items) fit the data best among the theoretically justified models tested. General factor was interpreted as math ability, foils and reals do not seem to be responded to using different processes.
14. & 15. School-level social pressure on domain achievements and school-level rule violation are related to overclaiming	*Failed:* No relation was found due to minimal across-school OCT scores variability.
16., 17. & 18. Socio-demographic correlates of overclaiming.	*Partially confirmed:* Male students yielded larger OCT bias than female students. Socio-economic status was only marginally related to OCT scores.
19. Exploring latent class structure of overclaiming: Can overclaiming class be identified?	*Confirmed:* Three latent classes were identified, one of which can be interpreted as an overclaiming class. It consists of 9% of the total sample size.

Table 48. Hypotheses tested and their verification.

8.2 Limitations

Study design

Nonetheless, the study presented above has its certain inherent limitations. First of all, it is a correlational study, so no evidence on the causal order of dependencies was provided. Although certain suggestions and conclusions on such relations can be inferred from the results presented in chapter 7 this is an obvious boundary to conclusions that can be taken from the information gathered. Such causal relations should be addressed in further experimental studies. It is worthy to note, that experimental manipulations in the field of OCT research are quite rare[128] (e.g. Atir et al., 2015; Atir et al., in preparation; Muller & Moshagen, 2018, 2019a, 2019b; Paulhus et al., 2003; Steger et al., 2020) and correlational designs constitute a predominant part of evidence.

Moreover, all presented analyses were cross-sectional ones, so no information on longitudinal effects of the presented associations can be formed. Such data is available for other response biases, especially important evidence was obtained with the use of longitudinal designs in RS research were large temporal stability of response patterns was found (He & van de Vijver, 2015b; Weijters et al., 2010). Wetzel, Luedtke, Zettler and Boenhke (2015) established that even up to 50% of RS variance can be attributed to trait-like source, which was stable for a period of over 8 years. The analysis comprised ERS and ARS analyses and introduced important information that RS may be of substantial interest too.

What are longitudinal patterns of OCT should be set up in future studies. The data at hand is scarce, but some interesting suggestions can be drawn from the research of Barber and others (2013) and especially from the evidence presented by von Hoyer, Pardi, Kammerer and Holtz (2019). In the latter study it was found that the tendency to overclaim increased during learning process (the participants had to learn about weather phenomena formation), which may point to memory bias driven by repetition effect (the overclaiming items were repeated several times during the task in few distinct time points). Moreover, overclaiming was related to overconfidence (r~ 0.40), especially in being confident of answer correctness when in fact an incorrect answer was provided. Negative relation between OCT scores and "metacognitive sensitivity", operationalised as difference in confidence between correct and incorrect answers, was also established. The relation between OCT scores and overconfidence effects disappeared after the learning process, but foils embracing increased as the result of the process. These results correspond to the finding of Atir and others (in preparation) who found evidence that foils are overclaimed to a larger degree when presented in a list of more familiar, in comparison to less familiar items. Both studies by Atir and colleagues, as well as von Hoyer and collaborators point to a significant role of memory processes in responding to foils. It is also worthy to link these results with the ones presented in subchapter 7.7 where foils were fitted in one factor with difficult items. Longitudinal studies would also help to answer questions about the test-retest reliability of OCT which may be low due to significant cognitive substrate in the task (on low intra-individual reliability of cognitive measures: Schubert & Frischkorn, 2020).

PISA 2012 characteristics

Furthermore, as the PISA 2012 cycle was one of the last ones to be held in paper-and-pencil mode no paradata data is available. As evidenced by e.g. Steger et al. (2020) such information may prove very useful in interpreting OCT scores. Future research endeavours should consider using RT, log-data or

[128] Of course if quite numerous instructed faking studies are to be treated separately (e.g. Bing et al., 2011).

mouse clicks. As PISA now also moved to computer-based mode such paradata is available for this ILSA too (though no PISA cycle since 2012 returned to the use of OCT again).

Moreover, also using the PISA data has its inherent limitations. First of all PISA is not the best for individual-level predictions and larger precision of estimates could be achieved if longer measurement tools (both cognitive tests and self-report questionnaires) could be used. PISA is also a low-stakes test which has its boundaries, e.g. stemming from problems with low motivation of certain participants. Furthermore, OCT in only one domain- academic knowledge in mathematics- was used in the study which confined analyses on the domain-specific vs. domain-general character of overclaiming. Further still, only data from one country was used which precludes any cross-culture and cross-country comparisons. Such studies, as evidenced by e.g. Fell and Koenig (2016; 2020) are very fruitful in interesting findings and certainly should be in the heart of researchers' interest in future studies (see Vonkova et al, 2018 for cross-country differences in the PISA 2012 OCT). However, this one-country approach (see also Yang et al., 2019) was necessary to build up evidence and theory before launching cross-country analyses. Many areas of OCT knowledge are still seriously under-researched and basic network of relations needs to be sketched before any confirmatory research will be possible. Finally, it is feasible to overcome PISA rotation design and impute the missing-by-design data that would rise the studies sample size and power (from around 3050 to 4600). However, most probably such step would not change the results obtained (cf. Borgonovi & Pokropek, 2019).

Methodological limitations

Most importantly, more sophisticated OCT measures that would better account for the ordinal character of Likert-type data are worthy of investigation. It is especially important if SDT measures should be still used in the response bias field. Their introduction to social sciences seems very promising, however, further research on their measurement and use are needed (especially for non-dichotomous data; also see Paulewicz & Blaut, 2020; Wright et al., 2009). More research is also needed to enhance interpretation of OCT scores, e.g. what is the meaning of correlation between the SDT scores, which is commonly obtained in empirical studies?

Moreover, many theoretically important variables were not measured in the PISA 2012 study or were measured only by non-established proxies (e.g. math anxiety). More studies using theoretically valid and methodologically sound measures of self-esteem, self-efficacy, math anxiety, locus of control, SDR scales, dark personality, personality (honesty-humility, openness to experience, competitive worldviews, conformity, unmitigated agency and all of these measures gauged on the facet/subscale level) are needed. Furthermore, not only zero-order correlations between these variables and OCT scores should be studied but also interactions between these variables need to be established, preferably in the structural equations modelling framework.

Another restraint of the analyses presented above is the fact that math ability was only measured by means of cognitive test. Such measure gives high-quality information on math achievement, but does not inform on other aspects of cognitive abilities, e.g. working memory, semantical fluency, etc. Combining such measures with OCT (e.g. Paulhus & Dubois, 2014) is warranted in order to broaden the knowledge about the relations between cognition and overclaiming.

8.3 Future directions

Theoretical advancements

Further research on the OCT relation to memory biases is needed. Especially warranted are research attempts aiming at understanding the role of repetition, false recognition and assimilation-contrast

effects. Also Roediger and McDermott (1995; 2000) offered suggestions for future studies on false memories attribution, semantic networks and false recognition effects (Is it familiar? Yes? Then I remember it!). Interesting knowledge on repetition effect comes from the Deese's (1959) list learning paradigm, where participants claim to recognise previously unstudied words if they are semantically related to the list studied (e.g. "needle" when studying medical terms list). Roediger and McDermott (1995), Roediger (1996) as well as Gallo et al. (1997) offer more information on this paradigm and the so-called "memory illusion". The relation between this paradigm and OCT seems obvious and worthy of further exploration. It could also bring some more light on the relation between OCT and hindsight bias (Muller, 2019; Williams, Paulhus & Nathanson, 2002) and foil plausibility (risk of confusion) manipulation (Calsyn et al., 2001; Franzen & Mader, 2019). Pennycook and Rand (2020) drive attention towards processing facilitation due to repetition and increased "truth-likeness" due to repetition (Whittlesea, 1993). This effect does not depend on neither explicit memory nor warnings against foils/fake news presence, moreover, it is not moderated by cognitive ability or style (De keersmacker et al., 2018). Repetition effect is thus steered by low-level cognition in which high-level cognition does not seem to intervene. Moreover, there is a communality between OCT accuracy and response time latencies in cognitive tasks (short RT=less accuracy) where intuitive, reflexive answer leads to incorrect response (exactly that was showed by Williams, Paulhus & Nathanson, 2002). Probably OCT may be (at least partially) also explained by similar processes. However, as indicated by Nichols and Loftus (2019), there is no evidence for a "false memory" trait as three different false-memory tasks do not share variance between them, yielding null or very small correlations. Hence, memory distortions are ubiquitous, but they do not seem to be driven by same mechanisms (see also Teovanović, Knezević & Stankov, 2015 for an account of generally low correlations between cognitive biases- the correlation coefficients are usually around 0.20). It is for future studies to investigate which memory biases underpin overclaiming and related effects.

Of course trivial memory slips, like forgetting or mistakes, are also possible in case of OCT (Bing et al., 2011), so it would be important to differentiate such overclaiming as a by-product of cognitive slips, fatigue, C/IER and RS from overclaiming driven by motivated biases (e.g. self-enhancement, SDR). To do this further research on how to generate and use C/IER indices is needed. Answering questions like: whether to generate them from the whole vector of responses the participant generated or only from the scale that is in the focus of interest (cf. Conijn et al., 2016)? What cut-off values to which conditions should be adopted? How to develop existing measures to the specific requirements of rating-scales? Similar questions appertain to the RS framework, e.g. what are the theoretical interpretations of different mapping matrices (Plieninger, 2020a)? How different RS should be modelled (Falk & Cai, 2016)? Do RS have substantial interpretations (Wetzel et al., 2015)? Why certain participants seem not to respond stylistically (Khorramdel et al., 2019)? How to differentiate a RS from a mere straight-lining as a result of satisficing?

Another problem waiting for further scrutiny is the reverse of overclaiming- underclaiming. Underclaimers are poorly researched but they were evidenced to exist- e.g. John and Robins (1994) found that 35% of the sample overestimated, but 15% underestimated. The distribution is thus tilted towards positivity but it does not cancel out the fact that underclaiming skills and knowledge is a certain mystery for scientific inquiry, especially as so far only minimal attention was devoted to studying underestimated self-evaluations. The results of Yang et al. (2019) show that a predominant group among low achieving students is not overclaiming- hence it is a challenge for Krueger-Dunning effect and similar explanations. These results, especially when linked with the ones obtained by Jerrim et al. (2019), where groups of higher socio-economic status overclaimed more, points to the existence of new, poorly researched groups: underclaimers, but also overclaimers of high cognitive abilities and

social status, which complement the most researched group of less cognitively skilled overclaimers (Bishop et al., 1986).

An interesting direction of future studies is to link the research on metacognitive abilities to assess one's self-knowledge (Tobias & Everson, 2002) and motivation to accurate self-knowledge (Brycz & Konarski, 2016; Brycz, Wyszomirska-Góra, Konarski & Wojciszke, 2018; Konarski & Brycz, 2017) with the positivity bias field. Such linkage would serve to further expand the nomological network of OCT, of course along with measuring other relevant variables, e.g. personality variables, creativity, etc. (see Goecke et al., 2020).

Methodological refinements

Apart from theoretical advancements, more research on design of overclaiming task is recommended. The method was so far proved to be robust to changes in design as in example Atir et al. (2015) showed that changing the main questions stem ("Do you know these things?", instead of "Are you familiar with…?") brought no significant difference into results. Similarly, including a "don't know" option also did not introduce any effects (Calsyn & Winter, 1999). Also foils-to-reals proportion is thought to have no effects on OCT measures (Paulhus et al., 2003), however both recent results presented by Atir et al. (in preparation) and knowledge on conflict resolution in cognitive conflict tasks (e.g. Stroop task, flanker task, etc.; Bugg & Crump, 2012) call for revisiting this question in a series of experimental tasks. Any influence of proportion of foils on OCT measures would point to a larger role of conflict resolution in this task that previously thought (cf. Robins & Beer, 2001). Interestingly, warning participants of foil presence seems to lower the bias, mostly by shifting participants strategy towards more conservative responding (Atir et al., 2015). More inter-domain comparisons would be also welcome, in order to test theories pointing to a motivated bias (e.g. SDR) as the main underpinning of overclaiming. Categories such as domain (item) importance, centrality, desirability and social prestige should be compared in experimental designs (Paulhus, 2011). Foils and reals construction rules is yet another topic worthy of more studies, if not for theoretical findings then definitely for guaranteeing methodological soundness of the research tools. Hargittai (2005) formulated a preliminary classification of foil-types but no differences between their different functioning was ever tested. This seems as an important research task to be realised, as OCT is a task that needs extensive piloting for both reals and foils. The latter can easily become obsolete, as can be seen by a look at the foils included in the OCT version by Randall and Fernandes (1991) or Philips and Clancy (1972). More inspiration for foils construction rules can be found in the works of Dubois (2015), Goecke et al. (2020) and Zimmerman et al. (1977; non-word construction rules).

Moreover, extending overclaiming tasks to other than rating-scales response format would be an interesting innovation. For example, OCT in multiple forced-choice (MFC) format would be very interesting (Wetzelf & Greiff, 2018), as well as embedding OCT logic into cognitive tests, e.g. through offering tests with no true answers but with a "don't know" option included (see a somewhat similar idea in Steger et al., 2020). Both response formats would predict lower OCT than rating-scales and, most importantly, would enable to restrict much of the (alleged) influence of RS and C/IER on OCT scores.

Additional sources of information that would be possible to correlate with OCT measures are also needed. Multi-method studies in which informant (other- vs. self-) ratings and more objective sources of information (cognitive tests, administrative data) along with qualitative studies (cognitive interviews) could bring more insight into respondents' perception of the OCT tasks. Obviously, combining self-reports with other measures such as paradata (RT, mouse clicks, navigation through web survey map, etc.), eye-tracker and both neurocognitive and psychophysiological measures.

Especially, the EEG use could yield interesting results on cognitive processes underpinning OCT responding. Psychophysiological correlates of mind-wandering, conflict resolution, semantical consolidation and word recognition all could immensely inform on OCT mechanisms (e.g. Arnau et al., 2020; Ryals, Yadon, Nomi & Cleary, 2011). Finally, it would be beneficial for the field if another ILSA or at least NLSA would include OCT into its questionnaire with a large-scale split ballot experiment being an especially interesting option.

Practical applications

The most direct and hoped-for effect of OCT refinement would be using it as a valid indicator of response biases in any research using self-reports, be it experimental small-sample research or a worldwide ILSA like PISA.

However, research on overclaiming and related effects could also have more direct, applicative deployments. For example, Lamba and Nityananda (2014) pointed into problems that self-deception can cause in financial and educational research. The researchers argued that promoting overconfident individuals[129] in societies may lead to catastrophic results such as economic crisis, wars and spreading of anti-communal values in societies (cf. Anderson et al., 2012). Stephens and Ohtsuka (2014) showed that cognitive, self-deceptive illusions (illusion of control) are one of the main characteristics of aggressive drivers. Such dependencies would be probably also found in other professions in risk of overconfidence (bankers, lawyers, medical doctors). Lamba and Nityananda (2014) concluded with proposing methods to identify self-deceived individuals as part of risk management in social institutions and companies, including schools, armies and banks (see also Gorlin & Otto, 2017). Overclaiming task is a potential option for such a measurement tool.

Methods identifying self-deceived and/or overconfident individuals could also help in educational process where they could be used to inform students whether their self-judgements, e.g. on school-related abilities, are adequate or not. Using such methods to teaching students to have more valid self-perceptions would be another worthy application (see Foster et al., 2017; Kim, Chiu & Zou, 2010).

Pennycook and Rand (2020) also called OCT "conceptually related" to reflexive responding, hence its relation to fake news and bullshit receptivity. Relation between OCT and overconfidence is also based on the tendency to be driven by gut-feeling instead of analytical thinking. Atir and colleagues (in preparation) also explained assimilation effect on the basis of automatic rather than deliberate thinking. Relation of OCT with accepting bullshit and propensity to fake news may also mean not a "memory bias" *sensu stricto*, but a general proclivity to claim familiarity with things or accepting things as making sense (despite they do not). This framework was called "reflexive open-mindedness" and is dedicated to research on being overly open to new ideas and deficits of criticism towards incoming information. Pennycook and Rand (2020) corroborated these ideas by a correlation between OCT and bullshit receptivity, fake news gullibility and less analytic thinking. Littrell, Risko and Fugelsang (2020) found a relation between overclaiming and persuasive (0.30) and evasive (0.19) bullshitting. It is interesting, that they definition of both types of "bullshitting" are in accord with personality correlates of overclaiming presented above. Moreover, individual differences in the trait-like construct called "an illusion of familiarity", an inherently subjective feeling, that interacts with prior experiences and knowledge of respondents were related to OCT bias (Paulhus, 2011; Whittlesea & Williams, 2000; 2001; see also "feeling of knowing" explanation by Clariana et al., 2016). Such effects can also explain the significance of foil plausibility manipulation done by Franzen and Mader (2019). Moreover, OCT potentially could be used in teaching participants how to discern true from faked news, which seems

[129] Due to self-deception, impression management or both.

as an important challenge before educational systems worldwide (see Nygren & Guath, 2019 for such attempts in the Swedish educational system). OCT could be sued to identify overclaimers which potentially are one of the groups in risk of increased fake news gullibility (cf. Pennycook & Rand, 2020).

8.4 Conclusion

The presented dissertation had three main aims. In order of importance:

- to assess utility of the overclaiming technique (OCT) to account for spurious variance in self-reports of skills in low-stakes, self-administered self-reports
- to investigate probable mechanisms of overclaiming by testing the proposed relationships in a cross-sectional, correlational design and by fitting latent variable models
- to expand the nomological network of overclaiming by presenting a wide gamut of correlations, both from individual- and school-level (where school was a basic clustering, as defined by the PISA design)

In order to formulate a set of verifiable hypotheses a thorough integration of the previous empirical and theoretical research was undertaken. Research fields like socially desirable responding, positivity bias, self-presentation, self-consciousness, self-knowledge, self-motives with a special attention on self-enhancement, response biases (e.g. response styles and careless responding) and self-reports validity were reviewed with the aim of finding relevant results to inform hypothesis generation in order to fulfil the three main aims. In order to test possible mechanisms of overclaiming three accounts were tested: a) overclaiming as a result of memory bias, b) overclaiming as a result of motivated bias, c) overclaiming as by-product of stylistic or careless responding. Moreover, confirmatory factor analysis and latent class analysis were fitted in order to explore latent structure of the PISA 2012 overclaiming scale. The analyses conducted resulted in following conclusions and findings:

1. Measure of OCT bias indeed can act as a suppressor, leading to enhanced predictive validity of self-report of skills on objective skills measurement. The established suppression model is of the reciprocal, not classical suppression due to non-zero raw correlation between OCT bias and objective skills measures. The magnitude of the predictive validity enhancement greatly depends on the OCT scoring system adopted with IRT and SDT measures leading to slightly different conclusions and need of different regression model specification. Special care is needed when interpreting OCT results as conclusions may differ greatly depending on the scoring system and the exact measures used.
2. Correlations between OCT measures, objective math ability, openness, perseverance and self-report indices of math ability (self-efficacy, experience with math problems scales) were tested in order to indirectly verify memory bias as mechanism of overclaiming. The obtained results pointed to positive relation between math ability and OCT accuracy and negative with OCT bias. Moreover, objectively-measured math ability was related to a higher degree with OCT scores than subjective reports on math ability. Furthermore, openness, perseverance and subjective math self-assessments were not related to OCT bias. This pattern of results seems to contradict the overgeneralisation account of memory bias hypothesis which states that overclaiming is a characteristic of more proficient respondents as they tend to overgeneralise the possessed knowledge accepting foils on the false premises that they resemble some chunks of veridical knowledge. In contrast, the results obtained appear to support the metacognitive account of memory bias hypothesis which states that overclaiming is characteristic for less proficient participants as a result of insufficient metacognitive control over the possessed knowledge. Finally, these results showed that OCT foils (untrue items used to elicit overclaiming) do not share much of the common variance with other self-report

scales. Is it due to the lack of shared substantial or method variance it is to be settled in future research projects.

3. In order to test relation between OCT and motivated bias of self-perception (positivity bias) a group of variables (e.g. declared math self-interest, motivation, social norms, learning ethic) was used as proxies of domain desirability which is a predictor of motivated bias. The obtained results pointed to a small in size, but robust positive relation between domain desirability and OCT bias pointing to the role of motivated bias of self-perception in emergence of overclaiming skills. However, the small size of the effect precludes from considering positivity bias as the main source of OCT variance. Moreover, perception of success control was related to OCT scores as previous theories suggested that less perceived control is related to more positivity bias that may serve as self-esteem protection from negative real-life outcomes. Indeed, external locus of control (belief that luck, coincidence or higher powers have greater influence over one's life than hard work, skills, etc.) was positively related to OCT bias, but the magnitude of this relation was again very small. Furthermore, motivated bias can be also fuelled by social expectations towards high achievements. An analysis of relations between social expectations, parental pressure and school policy regarding achievements (as reported by students and school principals) and OCT scores did not yield any significant results.
4. Motivated bias is related not only with exaggerating desirable virtues but also with concealing vices, e.g. social rule breaking. Hence, a group of variables indicating rule violation (truancy, school discipline) or social relations at school (e.g. teacher-student relations, sense of belonging to school) was related to OCT scores. Small evidence for relation between OCT scores and motivated bias was again found as OCT bias was related to claiming higher school belonging and better teacher-student relations. Both topics are typically perceived as socially desirable or even sensitive which corroborates such an interpretation of the patterns observed in this analysis.
5. School-level variables were also related to OCT scores. In this analysis variables like divergence between students' and school principals' opinions on school discipline, school-level social expectations for high math abilities and school-level rule violations were related to OCT measures. The results obtained no significant relations between any of the principal-reported variables and individual OCT levels. Moreover, the correlations found between school-level variables and OCT bias were very small in magnitude. Further scrutiny revealed that only a very limited portion of the OCT variance was attributable to the school-level of analysis. This pattern of results seem to suggest that social norms are either irrelevant to OCT scores or that these norms are non-variant across schools in Poland.
6. The analysis of overclaiming as a by-product of response styles and careless responding demonstrated that indeed more bias and lower OCT accuracy can be attributed to careless responding. However, the magnitude of this relation was only moderate. Moreover, some of the careless responding indices displayed internal problems which warrants further research in the field of inattentive responding to polish up careless responding indices methodology and interpretability. Self-report measures of inattentive responding were not related to OCT scores. Respondents' fatigue lowered OCT accuracy but did not influence OCT bias.
7. Response styles were measured using the IRTree framework and by fitting an IRT multidimensional model to the data. The results were a bit surprising at the first glance as extreme response style (preference for extreme response categories, e.g. "1" and "5" on a 1 to 5 rating scale) correlated negatively with OCT bias (and positively with OCT accuracy). It seems that at least in this overclaiming scale participants were very conservative when claiming familiarity with foils. This interpretation was corroborated by a positive relation between midpoint response style (preference for the middle response category on a response

scale) and OCT bias. This relation was again driven by strong respondents' preference for keeping to the negative side of the scale when assessing familiarity with OCT items. These results point that any relations between response styles and overclaiming may be very dependent upon the technical details of a given measurement tool as e.g. mapping matrices of response styles (way in which stylistic responses are coded), domain assessed, overclaiming scale characteristics (foils number, foils creation rules) and response scale features (e.g. number of categories, labelling). More knowledge on these issues need to be gathered in order to formulate any generalisable conclusions about the relation between response styles and overclaiming.

8. Multilevel confirmatory factor analysis was used to assess the PISA 2012 overclaiming scale's latent structure. This analysis was mainly performed in order to gauge whether reals and foils form the same or different latent factors. The conducted model fitting process yielded the bifactor solution with two specific factors: one for easy items and another for hard items and foils together. It seems that some foils were processed by participants more like reals while some reals were processed more like foils. Thus, it can be suggested that the ontic status of OCT items does not have decisive influence on how respondents respond to them. It appears that (subjective) item difficulty is much more important for the pattern of responses yielded. This analysis also corroborated that only a very low OCT variance proportion is attributable to the school-level.
9. It was found that boys overclaimed more than girls and that students coming from families of higher socio-economic status overclaimed less than students of lower status. The gender relation was independent from math ability, however, the socio-economic status did not keep its relation with OCT bias in a multiple regression equation when controlled for math ability. Various explanations for this pattern of results were suggested, e.g. indicating greater male susceptibility to memory biases, personality traits predicting overclaiming being more frequent among males and higher desirability of maths among male students in comparison to their female counterparts.
10. LCA examination yielded three latent classes, one of which was interpreted as "overclaiming" class due to a very positive self-report which was not warranted by objective abilities. Membership in this class was related to being male, of above-average socio-economic status and yielding very desirable profile also on other self-report scales. However, this group comprised only 9% of the total sample analysed.

Main contributions of this work entail thorough testing of suppression models and turning attention into different properties of various OCT scoring systems. Moreover, the dissertation gathered novel evidence on potential mechanisms of overclaiming under the memory bias, motivated bias and careless responding hypotheses. Furthermore, the results confirming low importance of school-level variables for overclaiming were gathered for the first time showing these relations in an in-depth way. Also for the first time a detailed study of overclaiming scale's latent structure was conducted with important suggestions for future studies regarding foils responding and proportion of variability accounted for different variance sources. Finally, relations between overclaiming and a large group of psychological and socio-demographical variables were gauged with some interesting finding regarding locus of control or gender role in explaining OCT scores. Furthermore, it was one of the very few positivity bias studies conducted on high school students in educational context and in other than Northern American culture which enlarged the generalization of the results gathered to date.

The study is also characterized by serious limitations, mainly by its cross-sectional, correlational design. Moreover, it analyses data from only one country and with the use of some non-standard measurement tools. However, these limitations are not that serious as large sample size and overall

high quality of the rich data gathered by the PISA 2012 study contributes to the sound quality of the evidence provided.

More experimental studies with the use of OCT are warranted especially to inform more on the mechanisms of overclaiming. However, apart from the research aimed at advancing theoretical knowledge a lot of studies targeted at expanding knowledge about the importance of technical details of overclaiming scales, especially character of foils (how are they constructed), proportion of foils to reals in a task or using other response formats than rating scales, are warranted due to large knowledge lacunas regarding even some most basic OCT characteristics. Moreover, response times analysis and measurement of psychophysiological data in overclaiming tasks is advocated as promising ways of advancing knowledge about this phenomenon.

Apart from its utility for the survey methodology field OCT can also potentially have practical applicability in other fields of social studies. For example, measurement tools able to validly gauge individual-level propensity towards overclaiming ca be used to enhance students' learning process, as they can inform students on the true level of their academic knowledge. Moreover, such tasks can help to monitor against overconfident decisions in fields where they could be especially costly (e.g. financial institutions, army, pilots, air control, etc.). Furthermore, overclaiming was evidenced to be a correlate of fake news gullibility. Hence, it is probable that OCT-like tasks can be used to train participants in order to improve their abilities to differentiate between true and faked information broadcasted.

In sum, overclaiming technique seems a valuable for task for both basic and applied social research and it is worthy of further studies regarding its methodological practice and theoretical interpretations.

9- REFERENCES

Abdi, H. (2007). Signal detection theory (SDT). *Encyclopedia of measurement and statistics*, 886-889.

Abele, A. E., & Brack, S. (2013). Preference for other persons' traits is dependent on the kind of social relationship. *Social Psychology*, 44, pp. 84-94. https://doi.org/10.1027/1864-9335/a000138.

Abele, A. E., Bruckmüller, S., & Wojciszke, B. (2014). You are so kind–and I am kind and smart: Actor–Observer Differences in the Interpretation of On-going Behavior. *Polish Psychological Bulletin*, *45*(4), 394-401.

Abele, A. E., & Wojciszke, B. (2007). Agency and communion from the perspective of self versus others. *Journal of Personality and Social Psychology, 93*(5), 751–763. https://doi.org/10.1037/0022-3514.93.5.751

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. [in:] *Advances in experimental social psychology* (Vol. 50, pp. 195-255). Academic Press.

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence*, *22*(2), 227–257.

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences*, *33*(4), 587–605. https://doi.org/10.1016/S0191-8869(01)00174-X

Ackerman, P. L., & Ellingsen, V. J. (2014). Vocabulary overclaiming—A complete approach: Ability, personality, self-concept correlates, and gender differences. *Intelligence, 46,* 216-227.

Adair, C. (2014). *Interventions for Addressing Faking on Personality Assessments for Employee Selection : A Meta- Analysis*. College of Science and Health Theses and Dissertations. 93, DePaul University.

Aichholzer, J. (2014). Random intercept EFA of personality scales. *Journal of Research in Personality*, *53*, 1–4. https://doi.org/10.1016/j.jrp.2014.07.001

Aichholzer, J. (2015). Controlling acquiescence bias in measurement invariance tests. *Psihologija*, *48*(4), 409–429. https://doi.org/10.2298/PSI1504409A

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In *Action control* (pp. 11-39). Springer, Berlin, Heidelberg.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Alicke, M. D. (1985). Global Self-Evaluation as Determined by the Desirability and Controllability of Trait Adjectives. *Journal of Personality and Social Psychology*, *49*(6), 1621–1630. https://doi.org/10.1037/0022-3514.49.6.1621

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. *The Self in Social Judgment*, *January 2005*, 85–106. https://doi.org/10.4324/9780203943250-11

Alicke, M. D., Sedikides, C., & Zhang, Y. (2019). The motivation to maintain favorable identities. *Self and Identity*, *8868*. https://doi.org/10.1080/15298868.2019.1640786

Alliger, G. M., Lilienfeld, S. O., & Mitchell, K. E. (1996). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science*, *7*(1), 32–39. https://doi.org/10.1111/j.1467-9280.1996.tb00663.x

Allison, S. T., Messick, D. M., & Goethals, G. R. (1989). On being better but not smarter than others: The Muhammad Ali effect. *Social Cognition*, *7*(3), 275–295.

Allport, G. W. (1928). A test for ascendance-submission. *Journal of Abnormal and Social Psychology*, *23*(2), 118–136. https://doi.org/10.1037/h0074218

Alschuler, A., Weinstein, G., Evans, J., Tamashiro, R., & Smith, W. (1977). Education for what? Measuring self-knowledge and levels of consciousness. *Simulation & Games*, *8*(1), 29–47.

Amati, F., Oh, H., Kwan, V. S. Y., Jordan, K., & Keenan, J. P. (2010). Overclaiming and the medial prefrontal cortex: A transcranial magnetic stimulation study. *Cognitive Neuroscience*, *1*(4), 268–276. https://doi.org/10.1080/17588928.2010.493971

Anderman, E. M., Griesinger, T., & Westerfield, G. (1998). Motivation and cheating during early adolescence. *Journal of Educational Psychology, 90*(1), 84–93. https://doi.org/10.1037/0022-0663.90.1.84

Anderson, C., Brion, S., Moore, D. A., & Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of Personality and Social Psychology*, *103*(4), 718–735. https://doi.org/10.1037/a0029395

Anderson, C., Srivastava, S., Beer, J. S., Spataro, S. E., & Chatman, J. A. (2006). Knowing your place: self-perceptions of status in face-to-face groups. *Journal of personality and social psychology*, *91*(6), 1094-1110.

Anderson, C. D., Warner, J. L., & Spencer, C. C. (1984). Inflation bias in self-assessment examinations: Implications for valid employee selection. *Journal of Applied Psychology*, *69*(4), 574–580. https://doi.org/10.1037/0021-9010.69.4.574

Anglim, J., Morse, G., De Vries, R. E., MacCann, C., & Marty, A. (2017). Comparing Job Applicants to Non-applicants Using an Item-level Bifactor Model on the HEXACO Personality Inventory. *European Journal of Personality*, *31*(6), 669–684. https://doi.org/10.1002/per.2120

Antonakis, J., House, R. J., & Simonton, D. K. (2017). Can super smart leaders suffer from too much of a good thing? The curvilinear effect of intelligence on perceived leadership behavior. *Journal of Applied Psychology*, *102*(7), 1003.

Arkin, R. M. (1981). Self-presentation styles. *Impression Management Theory and Social Psychological Research*, *311*, 334.

Arlin, M. (1976). Causal priority of social desirability over self-concept: a cross-lagged correlation analysis. *Journal of Personality and Social Psychology*, *33*(3), 267-272.

Arnau, S., Löffler, C., Rummel, J., Hagemann, D., Wascher, D. & Schubert, A.-L. (2020). Inter-trial alpha power indicates mind wandering. *Psychophysiology*. Advanced online publication. doi: 10.1111/psyp.13581

Asch, S. E. (1951). Effects of group pressure on the modification and distortion of judgments. [in:] H. Guetzkow (Ed.), *Groups, leadership, and men.* Pittsburgh, Pa.: Carnegie Press.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics—Theory and Methods*, *35*(3), 439-460.

Astra, R. L., & Singg, S. (2000). The role of self-esteem in affiliation. *The Journal of Psychology*, *134*(1), 15–22.

Atir, S., Rosenzweig, E., & Dunning, D. (2015). When Knowledge Knows No Bounds: Self-Perceived Expertise Predicts Claims of Impossible Knowledge. *Psychological Science*, *26*(8), 1295–1303. https://doi.org/10.1177/0956797615588195

Atir S., Rosenzweig, E., & Dunning D. A. (in preparation). Experts Know What They Don't Know: Genuine Knowledge Associated with Less Overclaiming. https://www.stavatir.com/papers

Atir S., Rosenzweig, E, & Dunning D. A. (in preparation). The Influence of Context on Overclaiming: When and Why Do People Claim to Know The Unknowable? https://www.stavatir.com/papers

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior research methods*, *45*(2), 527-535.

Avvisati, F. (2020). *Personal communication.*

Avvisati, F., & Keslair, F. (2014). "REPEST: Stata Module to Run Estimations with Weighted Replicate Samples and Plausible Values." *Statistical Software Components*, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s457918.html.

Bäckström, M. (2007). Higher-order factors in a five-factor personality inventory and its relation to social desirability. *European Journal of Psychological Assessment*, *23*(2), 63–70. https://doi.org/https://doi.org/10.1027/1015-5759.23.2.63

Bäckström, M., & Björklund, F. (2013). Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology*, *54*(2), 152–159. https://doi.org/10.1111/sjop.12015

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, *43*(3), 335–344. https://doi.org/10.1016/j.jrp.2008.12.013

Bäckström, M., Björklund, F., & Larsson, M. R. (2012). *Social desirability in personality assessment: Outline of a model to explain individual differences.* In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (p. 201–213). Oxford University Press.

Bäckström, M., Björklund, F., & Larsson, M. R. (2014). Criterion validity is maintained when items are evaluatively neutralized: Evidence from a full-scale five-factor model inventory. *European Journal of Personality*, *28*(6), 620-633.

Baczko-Dombi, A. (2017). Ucieczka od matematyki. Rekonstrukcja procesu w kontekście społecznego wizerunku przedmiotu. *Edukacja*, *140*(1), 39–54. https://doi.org/10.24131/3724.170103

Baer, A., Trumpeter, N. N., & Weathington, B. L. (2006). Gender differences in memory recall. *Modern Psychological Studies*, *12*(1), 3.

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M., & O'Donnell, O. (2008). Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, *17*(3), 351–375.

Bagozzi, R. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. In *Journal of Research in Personality* (Vol. 27, Issue 1, pp. 49–87). https://doi.org/10.1006/jrpe.1993.1005

Bakan, D. (1966). *The duality of human existence.* Chicago: Rand McNally.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* W H Freeman/Times Books/ Henry Holt & Co.

Barber, L. K., Barnes, C. M., & Carlson, K. D. (2013). Random and Systematic Error Effects of Insomnia on Survey Behavior. *Organizational Research Methods*, *16*(4), 616–649. https://doi.org/10.1177/1094428113493120

Bargh, J. A. (1990). Auto-motives: Preconscious determinants of thought and behavior. *Handbook of motivation and cognition*, *2*, 93-130.

Barrett, P. (2007). Structural equation modelling: adjudging model fit. *Personality and Individual Differences, 42(5),* 815–824. https://doi:10.1016/j.paid.2006.09.018

Barrios, V., Kwan, V. S. Y., Ganis, G., Gorman, J., Romanowski, J., & Keenan, J. P. (2008). Elucidating the neural correlates of egoistic and moralistic self-enhancement. *Consciousness and Cognition*, *17*(2), 451–456. https://doi.org/10.1016/j.concog.2008.03.006

Barzykowski, K., Leśniak, A., & Niedźwieńska, A. (2010). Rola wskazówek i przekonań w długotrwałej pamięci cen. *Roczniki Psychologiczne*, *13*(2), 125-144.

Baumeister, R. F., & Bushman, B. (2010). *Social psychology and human nature, brief version*. Nelson Education.

Becker, C. A. (1976). Allocation of attention during visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 556-566.

Beer, J. S. (2007). The default self: feeling good or being right?. *Trends in cognitive sciences*, *11*(5), 187-189.

Beer, J. S. (2014). Exaggerated Positivity in Self-Evaluation: A Social Neuroscience Approach to Reconciling the Role of Self-esteem Protection and Cognitive Bias. *Social and Personality Psychology Compass*, *8*(10), 583–594. https://doi.org/10.1111/spc3.12133

Beer, J. S., & Harris, M. A. (2019). The advantages and disadvantages of self-insight: New psychological and neural perspectives. In *Advances in Experimental Social Psychology* (1st ed., Vol. 60). Elsevier Inc. https://doi.org/10.1016/bs.aesp.2019.04.003

Beer, J. S., & Hughes, B. L. (2011). Self-enhancement: A social neuroscience perspective. [in:] M.D. Alicke and C. Sedikides (eds.) *Handbook of self-enhancement and self-protection.* New York/London: Guilford Press.

Beer, J. S., Lombardo, M. V., & Bhanji, J. P. (2010). Roles of Medial Prefrontal Cortex and Orbitofrontal Cortex in Self-evaluation. *Journal of Cognitive Neuroscience*, *22*(9), 2108–2119. https://doi.org/10.1162/jocn.2009.21359

Beer, J. S., Rigney, A. E., & Koski, J. E. (2018). Self-evaluation. In J. T. Wixted (Ed.), *Stevens' Handbook ofExperimental Psychology and Cognitive Neuroscience. Fourth Edition.* (pp. 1–30). John Wiley & Sons. https://doi.org/10.7551/mitpress/7458.003.0023

Bem, D. J. (1972). Self-perception theory. *Advances in experimental social psychology*, *6*(1), 1-62.

Bendixen, M. T. (1995). Compositional perceptual mapping using chi-squared trees analysis and correspondence analysis. *Journal of Marketing Management*, *11*(6), 571–581. https://doi.org/10.1080/0267257X.1995.9964368

Bensch, D. (2018). *The Nomological Network of Social Desirability and Faking : A Reappraisal*. Unpublished doctoral dissertation. Humboldt-Universität zu Berlin.

Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., & Ziegler, M. (2019). The Nature of Faking: A Homogeneous and Predictable Construct? *Psychological Assessment*, *31*(4), 532–544. https://doi.org/10.1037/pas0000619

Bensch, D., Paulhus, D. L., Stankov, L., & Ziegler, M. (2017/2019). Teasing Apart Overclaiming, Overconfidence, and Socially Desirable Responding. *Assessment*, *26*(3), 351–363. https://doi.org/10.1177/1073191117700268

Beretvas, S. N., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne social desirability scale. *Educational and Psychological Measurement*, *62*(4), 570–589. https://doi.org/10.1177/0013164402062004003

Bernreuter, R. G. (1933). The theory and construction of the personality inventory. *The Journal of Social Psychology*, *4*(4), 387-405.

Berry, C. M., Page, R. C., & Sackett, P. R. (2007). Effects of self-deceptive enhancement on personality-job performance relationships. *International Journal of Selection and Assessment*, *15*(1), 94–109. https://doi.org/10.1111/j.1468-2389.2007.00374.x

Bertsch, S., & Pesta, B. J. (2009). The Wonderlic Personnel Test and elementary cognitive tasks as predictors of religious sectarianism, scriptural acceptance and religious questioning. *Intelligence*, *37*(3), 231–237. https://doi.org/10.1016/j.intell.2008.10.003

Biderman, M. D., Nguyen, N. T., Cunningham, C. J. L., & Ghorbani, N. (2011). The ubiquity of common method variance: The case of the big five. *Journal of Research in Personality*, *45*(5), 417–429. https://doi.org/10.1016/j.jrp.2011.05.001

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural equation modeling*, *7*(4), 608-628.

Bing, M. N., Kluemper, D., Kristl Davison, H., Taylor, S., & Novicevic, M. (2011). Overclaiming as a measure of faking. *Organizational Behavior and Human Decision Processes*, *116*(1), 148–162. https://doi.org/10.1016/j.obhdp.2011.05.006

Birenbaum, M., & Montag, I. (1989). Style and substance in social desirability scales. *European Journal of Personality*, *3*(1), 47–59. https://doi.org/10.1002/per.2410030106

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Birney, R. C., Burdick, H., & Teevan, R. C. (1969). *Fear of failure*. Van Nostrand-Reinhold Company.

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1980). Pseudo-Opinions on Public Affairs. *Public Opinion Quarterly*, *44*(2), 198. https://doi.org/10.1086/268584

Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on Fictitious Issues: The Pressure to Answer Survey Questions. *Public Opinion Quarterly*, *50*(2), 240. https://doi.org/10.1086/268978

Black, M. (1982). The prevalence of humbug. *Philosophic Exchange*, *13*(1), 4.

Blasberg, S. A., Rogers, K. H., & Paulhus, D. L. (2014). The bidimensional impression management index (BIMI): Measuring agentic and communal forms of impression management. *Journal of Personality Assessment*, *96*(5), 523–531. https://doi.org/10.1080/00223891.2013.862252

Blasius, J., & Thiessen, V. (2012). *Assessing the Quality of Survey Data*. Newbury Park, CA: Sage.

Blasius, J., & Thiessen, V. (2015). Should we trust survey data? Assessing response simplification and data fabrication. *Social Science Research*, *52*(March), 479–493. https://doi.org/10.1016/j.ssresearch.2015.03.006

Blau, P. (1964). *Power and exchange in social life*. New York: John Wiley & Sons.

Block, J. (1965). *The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPI.* Appleton-Century-Crofts.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2014). Modeling Motivated Misreports to Sensitive Survey Questions. *Psychometrika*, *79*(3), 515–537. https://doi.org/10.1007/s11336-013-9390-9

Boeije, H., & Lensvelt-Mulders, G. (2002). Honest by chance: A qualitative interview study to clarify respondents'(non-) compliance with computer-assisted randomized response. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *75*(1), 24-39.

Boghossian, P. A. (1989). Content and self-knowledge. *Philosophical Topics*, *17*(1), 5-26.

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*(5), 335–352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*(5), 814–833. https://doi.org/10.1177/0013164410388411

Bong, M. (2008). Effects of parent-child relationships and classroom goal structures on motivation, help-seeking avoidance, and cheating. *The Journal of Experimental Education*, *76*(2), 191-217.

Bonneville-Roussy, A., Bouffard, T., & Vezeau, C. (2017). Trajectories of self-evaluation bias in primary and secondary school: Parental antecedents and academic consequences. *Journal of School Psychology*, *63*(August), 1–12. https://doi.org/10.1016/j.jsp.2017.02.002

Borgonovi, F., & Biecek, P. (2016). An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. *Learning and Individual Differences*, *49*, 128-137.

Borgonovi, F., & Pokropek, A. (2017). Mind that gap: The mediating role of intelligence and individuals' socio-economic status in explaining disparities in external political efficacy in 28 countries. *Intelligence*, *62*, 125–137. https://doi.org/10.1016/j.intell.2017.03.006

Borgonovi, F., & Pokropek, A. (2019). Seeing is believing: Task-exposure specificity and the development of mathematics self-efficacy evaluations. *Journal of Educational Psychology*, *111*(2), 268–283. https://doi.org/10.1037/edu0000280

Borkenau, P., & Amelang, M. (1985). The control of social desirability in personality inventories: A study using the principal-factor deletion technique. *Journal of Research in Personality*, *19*(1), 44-53.

Borkenau, P., & Zaltauskas, K. (2009). Effects of self-enhancement on agreement on personality profiles. *European Journal of Personality: Published for the European Association of Personality Psychology*, *23*(2), 107-123.

Bornholt, L. J., Goodnow, J. J., & Cooney, G. H. (1994). Influences of gender stereotypes on adolescents' perceptions of their own achievement. *American Educational Research Journal*, *31*(3), 675-692.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences*, *8*(12), 539–546.

Bowling, N. A., & Huang, J. L. (2018). Your Attention Please! Toward a Better Understanding of Research Participant Carelessness. *Applied Psychology*, *67*(2), 227–230. https://doi.org/10.1111/apps.12143

Bradley, J. V. (1981). Overconfidence in ignorant experts. *Bulletin of the Psychonomic Society*, *17*(2), 82–84.

Brandt, R. M. (1958). The accuracy of self-estimate: A measure of self-concept reality. *Genetic Psychology Monographs*.

Bredl, S., Storfinger, N., & Menold, N. (2011). A literature review of methods to detect fabricated survey data. *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, 3–24. http://www.econstor.eu/handle/10419/74449

Bredl, S., Winker, P., & Kötschau, K. (2012). A statistical approach to detect interviewer falsification of survey data. *Survey Methodology*, *38*(1), 1–10.

Brenner, P. S., & DeLamater, J. (2016). Lies, Damned Lies, and Survey Self-Reports? Identity as a Cause of Measurement Bias. *Social Psychology Quarterly*, *79*(4), 333–354. https://doi.org/10.1177/0190272516628298

Bridge, D. J. (2006). Memory & Cognition: What difference does gender make?. https://surface.syr.edu/cgi/viewcontent.cgi?article=1637&context=honors_capstone

Briggs, S. R., & Cheek, J. M. (1988). On the nature of self-monitoring: Problems with assessment, problems with validity. *Journal of Personality and Social Psychology*, *54*(4), 663-678.

Brouwer, D., Meijer, R. R., & Zevalkink, J. (2013). On the factor structure of the Beck Depression Inventory–II: G is the key. *Psychological Assessment*, 25(1), 136–145. https://doi.org/10.1037/a0029228

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460-502.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36-52.

Brown, G. T. L., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy and Practice*, *22*(4), 444–457. https://doi.org/10.1080/0969594X.2014.996523

Brummelman, E., Thomaes, S., Nelemans, S. A., De Castro, B. O., Overbeek, G., & Bushman, B. J. (2015). Origins of narcissism in children. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(12), 3659–3662. https://doi.org/10.1073/pnas.1420870112

Brutus, S., Gill, H., & Duniewicz, K. (2010). State of science in industrial and organizational psychology: A review of self-reported limitations. Personnel Psychology, 63, 907–936. doi:10.1111/j.1744-6570.2010 .01192.x

Brycz, H., & Konarski, R. (2016). Narzędzie do pomiaru metapoznawczego Ja: MJ-24. *Psychologia Społeczna,* 11(4), 509-526.

Brycz, H., Wyszomirska-Góra, M., Konarski, R., & Wojciszke, B. (2018). The metacognitive self fosters the drive for self-knowledge: The role of the metacognitive self in the motivation to search for diagnostic information about the self. *Polish Psychological Bulletin*, 49(1), 66–76.

Buckley, J. (2009). Cross-National Response Styles in International Educational Assessments : Evidence from PISA 2006. *Retrieved April*, *212*.

Bugg, J. M., & Crump, M. J. C. (2012). In support of a distinction between voluntary and stimulus-driven control: A review of the literature on proportion congruent effects. *Frontiers in Psychology*, *3*, 367.

Burns, G. N., & Christiansen, N. D. (2011). Methods of measuring faking behavior. *Human Performance*, *24*(4), 358–372. https://doi.org/10.1080/08959285.2011.597473

Burns, G. N., Fillipowski, J. N., Morris, M. B., & Shoda, E. A. (2015). Impact of electronic warnings on online personality scores and test-taker reactions in an applicant simulation. *Computers in Human Behavior*, *48*(July), 163–172. https://doi.org/10.1016/j.chb.2015.01.051

Burski, J., Chłoń-Domińczak, A., Palczyńska, M., Rynko, M., & Śpiewanowski, P. (2013). *Umiejętności Polaków – wyniki Międzynarodowego Badania Kompetencji Osób Dorosłych (PIAAC).* Instytut Badań Edukacyjnych.

Buss, A. H. (1980). *Self-consciousness and social anxiety*. Freeman.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, *65*(3), 245–281.

Cahill, D. P. (2015). Wishful Thinking, Fast and Slow. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences. https://dash.harvard.edu/bitstream/handle/1/17467495/CAHILL-DISSERTATION-2015.pdf?sequence=1

Cai, H., Wu, L., Shi, Y., Gu, R., & Sedikides, C. (2016). Self-enhancement among Westerners and Easterners: A cultural neuroscience approach. *Social Cognitive and Affective Neuroscience*, *11*(10), 1569–1578. https://doi.org/10.1093/scan/nsw072

Calsyn, R. J., & Winter, J. P. (1999). Understanding and controlling response bias in needs assessment studies. *Evaluation Review*, *23*(4), 399–417. https://doi.org/10.1177/0193841X9902300403

Calsyn, R. J., Kelemen, W. L., Jones, E. T., & Winter, J. P. (2001). Reducing overclaiming in needs assessment studies: An experimental comparison. *Evaluation Review*, *25*(6), 583–604. https://doi.org/10.1177/0193841X0102500601

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81-105.

Campbell J.D., Lavallee L.F. (1993) Who am I? The Role of Self-Concept Confusion in Understanding the Behavior of People with Low Self-Esteem. In: Baumeister R.F. (eds) *Self-Esteem. The Plenum Series in Social / Clinical Psychology*. Springer, Boston, MA

Campbell, J. D., & Tesser, A. (1983). Motivational interpretations of hindsight bias: An individual difference analysis. *Journal of Personality*, *51*(4), 605–620. https://doi.org/10.1111/j.1467-6494.1983.tb00868.x

Carden, S., Camper, T., & Holtzman, N. (2018). Cronbach's Alpha under Insufficient Effort Responding: An Analytic Approach. *Stats*, *2*(1), 1–14. https://doi.org/10.3390/stats2010001

Carretié, L., Hinojosa, J. A., Martín-Loeches, M., Mercado, F., & Tapia, M. (2004). Automatic attention to emotional stimuli: neural correlates. *Human Brain Mapping*, *22*(4), 290–299.

Cervellione, K. L., Lee, Y. S., & Bonanno, G. A. (2009). Rasch modeling of the self-deception scale of the balanced inventory of desirable responding. *Educational and Psychological Measurement*, *69*(3), 438–458. https://doi.org/10.1177/0013164408322020

Chachaj, A., Małyszczak, K., Poręba, R., Woźniak, D., Jabłońska, D., Cedzyński, Ł., ... & Szuba, A. (2006). Wybrane cechy osobowości osób z nadciśnieniem tętniczym. *Arterial Hypertension*, *10*(6), 532-537.

Chalmers, R., P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, 48(6), 1-29. doi: 10.18637/jss.v048.i06

Chambers, J. R. (2008). Explaining False Uniqueness: Why We are Both Better and Worse Than Others. *Social and Personality Psychology Compass*, *2*(2), 878–894. https://doi.org/10.1111/j.1751-9004.2008.00076.x

Chambers, J. R., & Suls, J. (2007). The role of egocentrism and focalism in the emotion intensity bias. *Journal of Experimental Social Psychology*, *43*(4), 618–625.

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin*, *130*(5), 813–838. https://doi.org/10.1037/0033-2909.130.5.813

Chance, Z., & Norton, M. I. (2015). The what and why of self-deception. *Current Opinion in Psychology*, *6*, 104–107. https://doi.org/10.1016/j.copsyc.2015.07.008

Chance, Z., Gino, F., Norton, M. I., & Ariely, D. (2015). The slow decay and quick revival of self-deception. *Frontiers in Psychology*, *6*(August), 1–6. https://doi.org/10.3389/fpsyg.2015.01075

Chance, Z., Norton, M. I., Gino, F., & Ariely, D. (2011). Temporal view of the costs and benefits of self-deception. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(SUPPL. 3), 15655–15659. https://doi.org/10.1073/pnas.1010658108

Christ, O., Schmid, K., Lolliot, S., Swart, H., Stolle, D., Tausch, N., Ramiah, A. Al, Wagner, U., Vertovec, S., & Hewstone, M. (2014). Contextual effect of positive intergroup contact on outgroup prejudice. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(11), 3996–4000. https://doi.org/10.1073/pnas.1320901111

Christensen, A. P., Cotter, K. N., & Silvia, P. J. (2019). Reopening Openness to Experience: A Network Analysis of Four Openness to Experience Inventories. *Journal of Personality Assessment*, *101*(6), 574–588. https://doi.org/10.1080/00223891.2018.1467428

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant assessment. *Human Performance*, *18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803

Christie, R., & Geis, F. L. (1970). *Machiavellianism*. Academic Press, Incorporated.

Chun, K. T., Campbell, J. B., & Yoo, J. H. (1974). Extreme response style in cross-cultural research: A reminder. *Journal of Cross-Cultural Psychology*, *5*(4), 465-480.

Chung, J., Schriber, R. A., & Robins, R. W. (2016). Positive Illusions in the Academic Context: A Longitudinal Study of Academic Self-Enhancement in College. *Personality and Social Psychology Bulletin*, *42*(10), 1384–1401. https://doi.org/10.1177/0146167216662866

Church, A. T., Katigbak, M. S., Del Prado, A. M., Ortiz, F. A., Mastor, K. A., Harumi, Y., ... & Miramontes, L. G. (2006). Implicit theories and self-perceptions of traitedness across cultures: Toward integration of cultural and trait psychology perspectives. *Journal of Cross-Cultural Psychology*, *37*(6), 694-716.

Cipora, K., Szczygiel, M., Willmes, K., & Nuerk, H. C. (2015). Math anxiety assessment with the Abbreviated Math Anxiety Scale: Applicability and usefulness: Insights from the polish adaptation. *Frontiers in Psychology*, *6*(NOV). https://doi.org/10.3389/fpsyg.2015.01833

Cipora, K., Willmes, K., Szwarc, A., & Nuerk, H.-C. (2018). Norms and validation of the online and paper-and-pencil versions of the Abbreviated Math Anxiety Scale (AMAS) for Polish adolescents and adults. *Journal of Numerical Cognition*, *3*(3), 667–693. https://doi.org/10.5964/jnc.v3i3.121

Cisłak, A. (2013). Effects of power on social perception. *Social Psychology*, 44, 138-146.

Clariana, M., Castelló, A., & Cladellas, R. (2016). Feeling of knowing and over-claiming in students from secondary school to university. *Learning and Individual Differences*, *49*, 421–427. https://doi.org/10.1016/j.lindif.2016.05.008

Clark, R. A., & Goldsmith, R. E. (2005). Market mavens: Psychological influences. *Psychology & Marketing*, *22*(4), 289-312.

Clark, S. L., Muthén, B., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and Strategies for Factor Mixture Analysis: An Example Concerning the Structure Underlying Psychological Disorders. *Structural Equation Modeling*, *20*(4), 681–703. https://doi.org/10.1080/10705511.2013.824786

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Coleman, N., & Mahaffey, T. (2000). Business student ethics: Selected predictors of attitudes toward cheating. *Teaching Business Ethics*, *4*(2), 121–136.

Colvin, C.R., Block, J. (1994). Do positive illusions foster mental health? An examination of the Taylor and Brown formulation. *Psychological Bulletin,* vol. 116, no. 1, 3-20.

Colvin, C. R., Block, J., & Funder, D. C. (1995). Overly positive self-evaluations and personality: negative implications for mental health. *Journal of Personality and Social Psychology*, *68*(6), 1152.

Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, *34*(1), 35–46.

Conger, A. J., & Jackson, D. N. (1972). Suppressor variables, prediction, and the interpretation of psychological relationships. *Educational and Psychological Measurement*, *32*(3), 579–599.

Conijn, J. M., Emons, W. H. M., & Sijtsma, K. (2014). Statistic lz-Based Person-Fit Methods for Noncognitive Multiscale Measures. *Applied Psychological Measurement*, *38*(2), 122–136. https://doi.org/10.1177/0146621613497568

Conijn, J. M., Emons, W. H. M., De Jong, K., & Sijtsma, K. (2015). Detecting and Explaining Aberrant Responding to the Outcome Questionnaire–45. *Assessment*, *22*(4), 513–524. https://doi.org/10.1177/1073191114560882

Conijn, J. M., Emons, W. H. M., van Assen, M. A. L. M., & Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research*, *46*(2), 365–388. https://doi.org/10.1080/00273171.2010.546733

Conijn, J. M., Sijtsma, K., & Emons, W. H. M. (2016). Identifying Person-Fit Latent Classes, and Explanation of Categorical and Continuous Person Misfit. *Applied Psychological Measurement*, *40*(2), 128–141. https://doi.org/10.1177/0146621615611164

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment*, *15*(1), 110-117.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire* (No. 63). Sage.

Cook, S. W., & Selltiz, C. (1964). A multiple-indicator approach to attitude measurement. *Psychological Bulletin*, *62*(1), 36–55. https://doi.org/10.1037/h0040289

Cooley, C. H. (1902). Looking-glass self. *The production of reality: Essays and readings on social interaction*, *6*.

Coopersmith, S. (1967). *The antecedents of self-esteem.* San Francisco: H. Freeman and Company.

Cornelis, I., Van Hiel, A., Roets, A., & Kossowska, M. (2009). Age differences in conservatism: Evidence on the mediating effects of personality and cognitive style. *Journal of Personality*, *77*(1), 51–88.

Cosentino, A. C., & Solano, A. C. (2008). Adaptación y validación Argentina de la marlowe-crowne social desirability scale. *Interdisciplinaria*, *25*(2), 197-216.

Couper, M. P. (2005). Technology trends in survey data collection. *Social Science Computer Review*, *23*(4), 486–501.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods and Research*, *40*(1), 169–193. https://doi.org/10.1177/0049124110390768

Crocker, J. (2002). The costs of seeking self–esteem. *Journal of Social Issues*, *58*(3), 597-615.

Crocker, J., & Park, L. E. (2003). *Seeking self-esteem: Construction, maintenance, and protection of self-worth.*

Crocker, J., & Park, L. E. (2004). The costly pursuit of self-esteem. *Psychological Bulletin*, *130*(3), 392–414. https://doi.org/10.1037/0033-2909.130.3.392

Cronbach, L. J. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology, 32*(7), 533–543. https://doi.org/10.1037/h0058518

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. https://doi.org/10.1177/001316444600600405

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*(1), 3–31. https://doi.org/10.1177/001316445001000101

Cross, P. (1977). Not can but will college teachers be improved? *New Directions for Higher Education*, 17, 1–15.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Cruyff, M. J. L. F., Böckenholt, U., & van der Heijden, P. G. M. (2016). The multidimensional randomized response design: Estimating different aspects of the same sensitive behavior. *Behavior Research Methods*, *48*(1), 390–399. https://doi.org/10.3758/s13428-015-0583-2

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Csikszentmihalyi, M., & Figurski, T. J. (1982). Self-awareness and aversive experience in everyday life. *Journal of personality*, *50*(1), 15-19.

Dale, J., & Weinberg, R. (1990). Burnout in sport: A review and critique. *Journal of Applied Sport Psychology*, *2*(1), 67–83.

Damarin, F., & Messick, S. (1965). Response Styles As Personality Variables: a Theoretical Integration of Multivariate Research1. *ETS Research Bulletin Series*, *1965*(1), i–116. https://doi.org/10.1002/j.2333-8504.1965.tb00967.x

Dashen, M. (2000). The effects of retention intervals on self-and proxy reports of purchases. *Memory*, *8*(3), 129-143.

Davidov, E., Muthen, B., & Schmidt, P. (2018). Measurement invariance. *Sociological Methods & Research*, *47*.

Davies, S. E., Connelly, B. S., Ones, D. S., & Birkland, A. S. (2015). The General Factor of Personality: The "Big One," a self-evaluative trait, or a methodological gnat that won't go away?. *Personality and Individual Differences*, *81*, 13-22.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17.

De Jong, M. G., Pieters, R., & Fox, J. P. (2010). Reducing social desirability bias through item randomized response: An application to measure underreported desires. *Journal of Marketing Research*, *47*(1), 14–27. https://doi.org/10.1509/jmkr.47.1.14

Deffler, S. A., Leary, M. R., & Hoyle, R. H. (2016). Knowing what you know: Intellectual humility and judgments of recognition memory. *Personality and Individual Differences*, *96*, 255–259. https://doi.org/10.1016/j.paid.2016.03.016

de-Magistris, T., & Pascucci, S. (2014). The effect of the solemn oath script in hypothetical choice experiment survey: A pilot study. *Economics Letters*, *123*(2), 252-255.

DeMaio, T. J. (1984). Social desirability and survey measurement: a review. [in:] Ch. Turner and Elizabeth Martin (Eds.), *Surveying Subjective Phenomena*, *2*, Russell Sage Foundation, New York, 257-281.

DeMars, C. E. (2013). A Tutorial on Interpreting Bifactor Model Scores. *International Journal of Testing*, *13*(4), 354–378. https://doi.org/10.1080/15305058.2013.799067

DeNisi, A. S., & Shaw, J. B. (1977). Investigation of the uses of self-reports of abilities. *Journal of Applied Psychology*, *62*(5), 641–644. https://doi.org/10.1037/0021-9010.62.5.641

Dennett, D. (1992). The Self as a Narrative Center of Gravity. [in:] D Dennett, F Kessel, P Cole, D Johnson (eds.). *Self and Consciousness: Multiple Perspectives. Hillsdale: Erlbaum*.

De Pascalis, V. (1993). Hemispheric asymmetry, personality and temperament. *Personality and Individual Differences*, *14*(6), 825-834.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, 70, 979–995

de Vries, R. E., Zettler, I., & Hilbig, B. E. (2014). Rethinking Trait Conceptions of Social Desirability Scales: Impression Management as an Expression of Honesty-Humility. *Assessment*, *21*(3), 286–299. https://doi.org/10.1177/1073191113504619

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of personality and social psychology*, *73*(6), 1246-1256.

Dijkstra, J. K., Lindenberg, S., & Veenstra, R. (2008). Beyond the class norm: Bullying behavior of popular adolescents and its relation to peer acceptance and rejection. *Journal of Abnormal Child Psychology*, *36*(8), 1289.

Dilchert, S. & Ones, D.S. (2012). Application of preventive strategies. [in:] Matthias Ziegler, Carolyn MacCann, Richard D. Roberts (Eds.), *New perspectives on faking in personality assessment,* Oxford University Press, 177-201.

Djikic, M., Chan, I., & Peterson, J. B. (2007). Reducing memory distortions in egoistic self-enhancers: Effects of indirect social facilitation. *Personality and Individual Differences*, *42*(4), 723–731. https://doi.org/10.1016/j.paid.2006.08.012

Djikic, M., Peterson, J. B., & Zelazo, P. D. (2005). Attentional biases and memory distortions in self-enhancers. *Personality and Individual Differences*, *38*(3), 559–568. https://doi.org/10.1016/j.paid.2004.05.010

Drwal, R. Ł., & Wilczyńska, J. T. (1980). Opracowanie kwestionariusza aprobaty społecznej [Elaboration of the social desirability questionnaire]. *Przegląd Psychologiczny, 23*(3), 569–583.

Dodou, D., & De Winter, J. C. F. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, *36*, 487–495. https://doi.org/10.1016/j.chb.2014.04.005

Donovan, W. L., Leavitt, L. A., & Walsh, R. O. (1990). Maternal self-efficacy: Illusory control and its effect on susceptibility to learned helplessness. *Child Development*, *61*(5), 1638–1647.

Dueber, D. M. (2017). *Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models.* https://dx.doi.org/10.13023/edp.tool.01 [Available at http://sites.education.uky.edu/apslab/resources/]

Dufner, M., Gebauer, J. E., Sedikides, C., & Denissen, J. J. A. (2019). Self-Enhancement and Psychological Adjustment: A Meta-Analytic Review. *Personality and Social Psychology Review*, *23*(1), 48–72. https://doi.org/10.1177/1088868318756467

Dunkel, C. S., van der Linden, D., Brown, N. A., & Mathes, E. W. (2016). Self-report based General Factor of Personality as socially-desirable responding, positive self-evaluation, and social-effectiveness. *Personality and Individual Differences*, *92*, 143–147. https://doi.org/10.1016/j.paid.2015.12.034

Dunlop, P. D., Bourdage, J. S., de Vries, R. E., Hilbig, B. E., Zettler, I., & Ludeke, S. G. (2017). Openness to (reporting) experiences that one never had: Overclaiming as an outcome of the knowledge accumulated through a proclivity for cognitive and aesthetic exploration. *Journal of Personality and Social Psychology*, *113*(5), 810–834. https://doi.org/10.1037/pspp0000110

Dunlop, P. D., Bourdage, J. S., de Vries, R. E., McNeill, I. M., Jorritsma, K., Orchard, M., Austen, T., Baines, T., & Choe, W. K. (2019). Liar! Liar! (When Stakes Are Higher): Understanding How the Overclaiming Technique Can Be Used to Measure Faking in Personnel Selection. *Journal of Applied Psychology*, *October*. https://doi.org/10.1037/apl0000463

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual Response Variability as an Indicator of Insufficient Effort Responding: Comparison to Other Indicators and Relationships with Individual Differences. *Journal of Business and Psychology*, *33*(1), 105–121. https://doi.org/10.1007/s10869-016-9479-0

Duval, S., & Wicklund, R. A. (1972). *A theory of objective self-awareness.* Academic Press.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, *16*(1), 1–23. https://doi.org/10.1207/S15327043HUP1601_1

Dwight, S. A., & Feigelson, M. E. (2000). A quantitative review of the effect of computerized testing on the measurement of social desirability. *Educational and Psychological Measurement*, *60*(3), 340–360. https://doi.org/10.1177/00131640021970583

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, *37*(2), 90–93. https://doi.org/10.1037/h0058073

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. Dryden Press.

Edwards, A. L., Diers, C. J., & Walker, J. N. (1962). Response sets and factor loadings on sixty-one personality scales. *Journal of Applied Psychology*, *46*(3), 220–225. https://doi.org/10.1037/h0040280

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356.

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education*, *28*(2), 497–510.

Emons, W. H. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*(3), 224-247.

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from patterns of summed polytomous item scores. *Applied Psychological Measurement*, *33*(8), 599–619. https://doi.org/10.1177/0146621609334378

Ennis, P. H. (1965). *Adult book reading in the United States: A preliminary report* (Issue 105). National Opinion Research Center.

Epley, N., & Dunning, D. (2000). Feeling" holier than thou": are self-serving assessments produced by errors in self-or social prediction?. *Journal of personality and social psychology*, *79*(6), 861-875.

Epley, N., & Whitchurch, E. (2008). Mirror, mirror on the wall: Enhancement in self-recognition. *Personality and Social Psychology Bulletin*, *34*(9), 1159–1170. https://doi.org/10.1177/0146167208318601

Eysenck, S. B. G., & Eysenck, H. J. (1968). The measurement of psychoticism: a study of factor stability and reliability. *British Journal of Social and Clinical Psychology*, *7*(4), 286-294.

Eysenck, S. B. G., Eysenck, H. J., & Shaw, L. (1974). The modification of personality and lie scale scores by special “honesty” instructions. *British Journal of Social and Clinical Psychology*, *13*(1), 41–50. https://doi.org/10.1111/j.2044-8260.1974.tb00876.x

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328–347. https://doi.org/10.1037/met0000059

Farrow, T. F. D., Burgess, J., Wilkinson, I. D., & Hunter, M. D. (2015). Neural correlates of self-deception and impression-management. *Neuropsychologia*, *67*, 159–174. https://doi.org/10.1016/j.neuropsychologia.2014.12.016

Fazekas, P., & Overgaard, M. (2018). *Perceptual consciousness and cognitive access: an introduction*. The Royal Society.

Federowicz, M. (red.). (2012). *Program Międzynarodowej Oceny Umiejętności Uczniów. Programme For International Student Assessment. Wyniki badania 2012 w Polsce*. Ministerstwo Edukacji Narodowej. Warszawa.

Feeney, J. R., & Goffin, R. D. (2015). The Overclaiming Questionnaire: A good way to measure faking? *Personality and Individual Differences*, *82*(August), 248–252. https://doi.org/10.1016/j.paid.2015.03.038

Fell, C. B., & König, C. J. (2016). Cross-Cultural Differences in Applicant Faking on Personality Tests: A 43-Nation Study. *Applied Psychology*, *65*(4), 671–717. https://doi.org/10.1111/apps.12078

Fell, C. B., & König, C. J. (2020). Examining Cross-Cultural Differences in Academic Faking in 41 Nations. *Applied Psychology*, *69*(2), 444–478. https://doi.org/10.1111/apps.12178

Fell, C. B., König, C. J., Jung, S., Sorg, D., & Ziegler, M. (2019). Are country level prevalences of rule violations associated with knowledge overclaiming among students? *International Journal of Psychology*, *54*(1), 17–22. https://doi.org/10.1002/ijop.12441

Felson, R. B. (1981). Self-and reflected appraisal among football players: A test of the Meadian hypothesis. *Social Psychology Quarterly*, 116–126.

Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, *43*(4), 522–527. https://doi.org/10.1037/h0076760

Fernbach, P. M., Hagmayer, Y., & Sloman, S. A. (2014). Effort denial in self-deception. *Organizational Behavior and Human Decision Processes*, *123*(1), 1–8.

Ferrando, P. J. (2005). Factor analytic procedures for assessing social desirability in binary items. *Multivariate Behavioral Research*, 40(3), 331-349.

Ferraro, P. J. (2010). Know thyself: Competence and self-awareness. *Atlantic Economic Journal*, *38*(2), 183–196. https://doi.org/10.1007/s11293-010-9226-2

Festinger, L. (1954). A theory of social comparison processes. *Human relations*, *7*(2), 117-140.

Fisher, R. J. (1993). Social desirability indirect questioning. *Journal of Consumer Research*, *20*(September 1993), 303–315. https://doi.org/10.1086/209351

Flagan, T., & Beer, J. S. (2013). Three ways in which midline regions contribute to self-evaluation. *Frontiers in Human Neuroscience*, *7*(JUL), 1–12. https://doi.org/10.3389/fnhum.2013.00450

Fleeson, W., & Wilt, J. (2010). The relevance of Big Five trait content in behavior to subjective authenticity: Do high levels of within-person behavioral variability undermine or enable authenticity achievement?. *Journal of Personality*, *78*(4), 1353-1382.

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive responding in MTurk and other online samples. *Industrial and Organizational Psychology*, *8*(2), 196-202.

Försterling, F., & Morgenstern, M. (2002). Accuracy of self-assessment and task performance: Does it pay to know the truth? *Journal of Educational Psychology, 94*(3), 576–585. https://doi.org/10.1037/0022-0663.94.3.576

Foster, J. D., Campbell, W. K., & Twenge, J. M. (2003). Individual differences in narcissism: Inflated self-views across the lifespan and around the world. *Journal of Research in Personality*, *37*(6), 469-486.

Foster, N. L., Was, C. A., Dunlosky, J., & Isaacson, R. M. (2017). Even after thirteen class exams, students are still overconfident: the role of memory for past exam performance in student predictions. *Metacognition and Learning*, *12*(1). https://doi.org/10.1007/s11409-016-9158-6

Fotios, S., & Gibbons, R. (2018). Road lighting research for drivers and pedestrians: The basis of luminance and illuminance recommendations. *Lighting Research & Technology*, *50*(1), 154-186.

Fraboni, M., & Cooper, D. (1989). Further validation of three short forms of the Marlowe-Crowne Scale of Social Desirability. *Psychological Reports*, *65*(2), 595-600.

Frank, M. R., Cebrian, M., Pickard, G., & Rahwan, I. (2017). Validating Bayesian truth serum in large-scale online human experiments. *PLoS ONE*, *12*(5), 1–13. https://doi.org/10.1371/journal.pone.0177385

Frankel, J., & Sharp, M. L. (1981). Measurement of Respondent Burden: Summary of Study Design and Early Findings. *Statistical Reporter*, *4*, 105–112. https://files.eric.ed.gov/fulltext/ED198179.pdf

Frankfurt, H. (1986). On bullshit. *Raritan Quarterly Review,* 6, 81–100.

Frankfurt, H. G. On bullshit. 2005. Princeton University Press, Princeton, NJ.

Franzen, A., & Mader, S. (2019). Do phantom questions measure social desirability? *Methods, Data, Analyses*, *13*(1), 37–57. https://doi.org/10.12758/mda.2019.01

Frenkel-Brunswik, E. (1939). Mechanisms of Self-Deception. *Journal of Social Psychology*, *10*(3), 407–420. https://doi.org/10.1080/00224545.1939.9713377

Freud, S. (1938/1941). Splitting of the ego in the defensive process. *International Journal of Psychoanalysis,* 22, 65–68.

Fritz, M. F. (1927). Guessing in a true-false test. *Journal of Educational Psychology*, *18*(8), 558–561. https://doi.org/10.1037/h0074440

Fronczyk, K. (2014). The identification of random or careless responding in questionnaires: The example of the NEO-FFI. *Roczniki Psychologiczne*, *17*(2), 439–473.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, *7*(3), 385–400. https://doi.org/10.1016/0191-8869(86)90014-0

Furnham, A. (1990). Faking personality questionnaires: Fabricating different profiles for different purposes. *Current Psychology*, *9*(1), 46–55. https://doi.org/10.1007/BF02686767

Furnham, A., & Chamorro-Premuzic, T. (2006). Personality, intelligence and general knowledge. *Learning and Individual Differences*, *16*(1), 79–90. https://doi.org/10.1016/j.lindif.2005.07.002

Furnham, A., & Henderson, M. (1982). The good, the bad and the mad: Response bias in self-report measures. *Personality and Individual Differences*, *3*(3), 311–320. https://doi.org/10.1016/0191-8869(82)90051-4

Gabriel, M. T., Critelli, J. W., & Ee, J. S. (1994). Narcissistic Illusions in Self-Evaluations of Intelligence and Attractiveness. *Journal of Personality*, *62*(1), 143–155. https://doi.org/10.1111/j.1467-6494.1994.tb00798.x

Gadzella, B. M., Cochran, S. W., Parham, L., & Fournet, G. P. (1976). Accuracy and differences among students in their predictions of semester achievement. *The Journal of Educational Research*, *70*(2), 75–81.

Gage, N. L., Leavitt, G. S., & Stone, G. C. (1957). The psychological meaning of acquiescence set for authoritarianism. *The Journal of Abnormal and Social Psychology, 55*(1), 98–103. https://doi.org/10.1037/h0047556

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, *72*(5), 892-913.

Gallo, D. A., Roberts, M. J., & Seamon, J. G. (1997). Remembering words not presented in lists: Can we avoid creating false memories? *Psychonomic Bulletin and Review*, *4*(2), 271–276. https://doi.org/10.3758/BF03209405

Ganassali, S. (2008). The influence of the design of web survey questionnaires on the quality of responses. *Survey Research Methods*, *2*(1), 21–32. https://doi.org/10.18148/srm/2008.v2i1.598

Gangestad, S. W., & Snyder, M. (2000). Self-monitoring: Appraisal and reappraisal. *Psychological Bulletin*, *126*(4), 530–555. https://doi.org/10.1037/0033-2909.126.4.530

Ganster, D. C., Hennessey, H. W., & Luthans, F. (1983). Social Desirability Response Effects: Three Alternative Models. *Academy of Management Journal*, *26*(2), 321–331. https://doi.org/10.5465/255979

Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating beliefs under perceived threat. *Journal of Neuroscience*, *38*(36), 7901–7911. https://doi.org/10.1523/JNEUROSCI.0716-18.2018

Gebauer, J. E., Sedikides, C., & Schrade, A. (2017). Christian self-enhancement. *Journal of Personality and Social Psychology*, *113*(5), 786–809. https://doi.org/10.1037/pspp0000140

Gebauer, J. E., Sedikides, C., Verplanken, B., & Maio, G. R. (2012). Communal Narcissism. *Journal of Personality and Social Psychology*, *103*(5), 854–878. https://doi.org/10.1037/a0029629

Gebauer, J. E., Wagner, J., Sedikides, C., & Neberich, W. (2013). Agency-communion and self-esteem relations are moderated by culture, religiosity, age, and sex: Evidence for the "self-centrality breeds self-enhancement" principle. *Journal of Personality*, *81*(3), 261–275. https://doi.org/10.1111/j.1467-6494.2012.00807.x

Gergen, J. K., & Gergen, M. M. (1980). *Causal attribution in the context of social explanation. In Gorlitz, D.(Ed.) Perspectives on attribution research and theory*.

Gesiarz, F., Cahill, D., & Sharot, T. (2019). Evidence accumulation is biased by motivation: A computational account. *PLoS Computational Biology*, *15*(6), 1–15. https://doi.org/10.1371/journal.pcbi.1007089

Gibby, R. E., & Zickar, M. J. (2008). A HISTORY OF THE EARLY DAYS OF PERSONALITY TESTING IN AMERICAN INDUSTRY: An Obsession With Adjustment. *History of Psychology*, *11*(3), 164–184. https://doi.org/10.1037/a0013041

Gibson, A. M., & Bowling, N. A. (2019). The Effects of Questionnaire Length and Behavioral Consequences on Careless Responding. *European Journal of Psychological Assessment*, 1–11. https://doi.org/10.1027/1015-5759/a000526

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence*, *80*(December 2019). https://doi.org/10.1016/j.intell.2020.101449

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development,* Cambridge, MA: Harvard University Press

Giromini, L., Viglione, D. J., Pignolo, C., & Zennaro, A. (2019). An Inventory of Problems–29 Sensitivity Study Investigating Feigning of Four Different Symptom Presentations Via Malingering Experimental Paradigm. *Journal of personality assessment*, 1-10.

Gnambs, T., & Kaspar, K. (2017). Socially Desirable Responding in Web-Based Questionnaires: A Meta-Analytic Review of the Candor Hypothesis. *Assessment*, *24*(6), 746–762. https://doi.org/10.1177/1073191115624547

Goecke, B., Weiss, S., Steger, D., Schroeders, U., & Wilhelm, O. (2020). Testing competing claims about overclaiming. *Intelligence, 81*, 101470.

Goffman, E. (1959). *The presentation of self in everyday life.* Anchor.

Gollwitzer, P. M. (1986). Striving for specific identities: The social reality of self-symbolizing. [in:] Roy F. Baumeister, *Public self and private self* (pp. 143-159). Springer, New York, NY.

Gordon, C. (1968). Self-conceptions; Configurations of content. *The Self in Social Interaction: Classic and Contemporary Perspectives*, *1*, 115–136.

Gordon, L. V. (1951). Validities of the forced-choice and questionnaire methods of personality measurement. *Journal of Applied Psychology*, *35*(6), 407–412. https://doi.org/10.1037/h0058853

Gordon, Ch., & Gergen, K.J. (Eds.) (1968). *The Self in Social Interaction. Vol. I: Classic and Contemporary Perspectives*. New York: John Wiley & Sons.

Gorlin, E. I., & Otto, M. W. (2017). *Truth matters: Cognitive integrity as an intervention for self-deception*. https://psyarxiv.com/72h65/download?format=pdf

Gramzow, R. H., & Willard, G. (2006). Exaggerating current and past performance: Motivated self-enhancement versus reconstructive memory. *Personality and Social Psychology Bulletin*, *32*(8), 1114–1125.

Gramzow, R. H., Elliot, A. J., Asher, E., & McGregor, H. A. (2003). Self-evaluation bias and academic performance: Some ways and some reasons why. *Journal of Research in Personality*, *37*(2), 41–61. https://doi.org/10.1016/S0092-6566(02)00535-4

Grau, I., Ebbeler, C., & Banse, R. (2019). Cultural Differences in Careless Responding. *Journal of Cross-Cultural Psychology*, *50*(3), 336–357. https://doi.org/10.1177/0022022119827379

Greenwald, A. (1985). Totalitarian egos in the personalities of democratic leaders. Symposium Paper. *International Society of Political Psychology Annual Meeting, Washington, DC, June*, *20*.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist*, *35*(7), 603–618. https://doi.org/10.1037/0003-066X.35.7.603

Greenwald, A. G. (1997). Self-knowledge and self-deception: Further consideration BT - The mythomanias: The nature of deception and self-deception. *The Mythomanias: The Nature of Deception and Self-Deception*, *3*, 51–71. https://doi.org/10.1165/rcmb.2011-0289OC

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). *Journal of Personality and Social Psychology,* Vol 74(6), 1464-1480.

Gregg, A. P. (2007). When vying reveals lying: The timed antagonistic response alethiometer. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *21*(5), 621-647.

Gregg, A. P., & Klymowsky, J. (2013). The implicit association test in market research: potentials and pitfalls. *Psychology & Marketing*, *30*(7), 588-601.

Gresham, F. M., Lane, K. L., MacMillan, D. L., Bocian, K. M., & Ward, S. L. (2000). Effects of positive and negative illusory biases: Comparisons across social and academic self-concept domains. *Journal of School Psychology*, *38*(2), 151–175.

Griffith, R. L., & Peterson, M. H. (2008). The Failure of Social Desirability Measures to Capture Applicant Faking Behavior. *Industrial and Organizational Psychology*, *1*(3), 308–311. https://doi.org/10.1111/j.1754-9434.2008.00053.x

Griffith, R. L., & Peterson, M. H. (2011). One piece at a time: The puzzle of applicant faking and a call for theory. *Human Performance*, *24*(4), 291–301. https://doi.org/10.1080/08959285.2011.597474

Griffith, R. L., Malm, T., English, A., Yoshita, Y., & Gujar, A. (2006). Applicant faking behavior: Teasing apart the influence of situational variance, cognitive biases, and individual differences. *A Closer Examination of Applicant Faking Behavior*, 151–178.

Griffith, R. L., & McDaniel, M. (2006). The nature of deception and applicant faking behavior. [in:] Richard L. Griffith and Mitchell H. Peterson, *A closer examination of applicant faking behavior*, IAP, Greenwich, Connecticut, 1-19.

Grosz, M. P., Lösch, T., & Back, M. D. (2017). The narcissism-overclaiming link revisited. *Journal of Research in Personality*, *70*, 134–138. https://doi.org/10.1016/j.jrp.2017.05.006

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: Wiley.

Groves, R. M., Fowler, F.J., Couper, M. P., Lepkowski, J. M., Singer, E. and Tourangeau, R. (2009). *Survey Methodology*. New York: Wiley

Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public opinion quarterly*, *74*(5), 849-879.

Grudniewska, M., & Kondratek, B. (2012). Zróżnicowane funkcjonowanie zadań w egzaminach zewnętrznych w zależności od płci na przykładzie części matematyczno-przyrodniczej egzaminu gimnazjalnego. *XVIII Konferencja Diagnostyki Edukacyjnej, Wrocław*. http://www.ptde.org/pluginfile.php/750/mod_page/content/4/Archiwum/XVIII_KDE/XVIII%20KDE%20-%20referaty/Grudniewska%2CKondratek.pdf

Grzegorczyk, A. (1978). Refleksje o psychologicznej koncepcji człowieka. *Teksty: Teoria Literatury, Krytyka, Interpretacja*, *3 (39)*, 9–36.

Guadagno, R. E., & Cialdini, R. B. (2007). Gender differences in impression management in organizations: A qualitative review. *Sex Roles*, *56*(7–8), 483–494.

Gummer, T., Roßmann, J., & Silber, H. (2018). Using Instructed Response Items as Attention Checks in Web Surveys: Properties and Implementation. *Sociological Methods and Research*. https://doi.org/10.1177/0049124118769083

Guo, S., Liu, Y., Wang, Y., Li, L. M. W., & Gao, D. (2019). Impression management in predicting social stress and adaptive work behaviors. *International Journal of Stress Management*. https://doi.org/10.1037/str0000143

Gur, R. C., & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, *37*(2), 147–169. https://doi.org/10.1037/0022-3514.37.2.147

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160-170.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research*, *52*(1), 123–142.

Hargittai, E. (2005). Survey measures of web-oriented digital literacy. *Social Science Computer Review*, *23*(3), 371–379. https://doi.org/10.1177/0894439305275911

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of educational psychology*, *56*(4), 208-216.

Harter, S. (2012). Self-perception profile for adolescents: Manual and questionnaires. *Denver, CO: Univeristy of Denver, Department of Psychology*.

Hartshorne, H., & May, M. A. (1928/1930). *Studies in the nature of character*, 1-3.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota multiphasic personality inventory* (Rev. ed., 2nd printing). University of Minnesota Press.

Hauser, P. (1969). Comments on Coleman's Paper. [In:] Robert Biersted (eds.) *Design for Sociology: Scope, Objectives and Methods*. Philadelphia: American Academy of Political and Social Science.

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three–Testing the theory in structural equation models! *Personality and Individual Differences, 42(5),* 841-850.

He, J. C., & Côté, S. (2019). Self-insight into emotional and cognitive abilities is not related to higher adjustment. *Nature Human Behaviour*, *3*(8), 867–884. https://doi.org/10.1038/s41562-019-0644-0

He, J., & Van de Vijver, F. (2016). Correcting for Scale Usage Differences among Latin American Countries, Portugal, and Spain in PISA. *RELIEVE - Revista Electronica de Investigacion y Evaluacion Educativa*, *22*(1). https://doi.org/10.7203/relieve.22.1.8282

He, J., & van de Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences*, *55*(7), 794–800. https://doi.org/10.1016/j.paid.2013.06.017

He, J., & Van De Vijver, F. J. R. (2015a). Effects of a general response style on cross-cultural comparisons: Evidence from the teaching and learning international survey. *Public Opinion Quarterly*, *79*(S1), 267–290. https://doi.org/10.1093/poq/nfv006

He, J., & van de Vijver, F. J. R. (2015b). Self-presentation styles in self-reports: Linking the general factors of response styles, personality traits, and values in a longitudinal study. *Personality and Individual Differences*, *81*, 129–134. https://doi.org/10.1016/j.paid.2014.09.009

Heggestad, E. D. (2012). *A conceptual representation of faking: Putting the horse back in front of the cart.* In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (p. 87–101). Oxford University Press.

Heine, S. J., Lehman, D. R., Markus, H. R., & Kitayama, S. (1999). Is there a universal need for positive self-regard? *Psychological Review*, *106*(4), 766.

Helgeson, V. S. (1994). Relation of agency and communion to well-being: Evidence and potential explanations. *Psychological Bulletin, 116*(3), 412–428. https://doi.org/10.1037/0033-2909.116.3.412

Helgeson, V. S., & Fritz, H. L. (1999). Unmitigated agency and unmitigated communion: Distinctions from agency and communion. *Journal of Research in Personality*, *33*(2), 131–158.

Henry, K. L., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling*, 17(2), 193-215.

Henschel, S., & Roick, T. (2017). Relationships of mathematics performances, control and value beliefs with cognitive and affective math anxiety. *Learning and Individual Differences,* 55, 97-107. https://doi.org/10.1016/j.lindif.2017.03.009

Hepper, E. G., Gramzow, R. H., & Sedikides, C. (2010). Individual Differences in Self-Enhancement and Self-Protection Strategies: An Integrative Analysis. *Journal of Personality*, *78*(2), 781–814. https://doi.org/10.1111/j.1467-6494.2010.00633.x

Hernandez, R., Kershaw, K. N., Siddique, J., Boehm, J. K., Kubzansky, L. D., Diez-Roux, A., Ning, H., & Lloyd-Jones, D. M. (2015). Optimism and Cardiovascular Health: Multi-Ethnic Study of Atherosclerosis (MESA). *Health Behavior and Policy Review*, *2*(1), 62–73. https://doi.org/10.14485/hbpr.2.1.6

Herr, P. M., Sherman, S. J., & Fazio, R. H. (1983). On the consequences of priming: Assimilation and contrast effects. *Journal of experimental social psychology*, *19*(4), 323-340.

Herzog, A. R., & Bachman, J. G. (1981). Effects of Questionnaire Length on Response Quality. *Public Opinion Quarterly*, *45*(4), 549. https://doi.org/10.1086/268687

Higgins, E. T. (1987). Self-discrepancy: a theory relating self and affect. *Psychological Review*, *94*(3), 319-340.

Hilgard, E. R. (1949). Human motives and the concept of the self. *American Psychologist*, *4*(9), 374.

Hill, T., Smith, N. D., & Lewicki, P. (1989). The development of self-image bias: A real-world demonstration. *Personality and Social Psychology Bulletin*, *15*(2), 205-211.

Hipsz, N. (2014). *Studium nad efektem społecznej poprawności w wywiadzie kwestionariuszowym. Randomized response technique - rozumienie i efektywność*. Unpublished doctoral thesis, University of Warsaw, Warsaw, Poland.

Hochstim, J. R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American statistical Association*, *62*(319), 976-989.

Hogan, R. (1983). A socioanalytic theory of personality. In *Nebraska symposium on motivation*. University of Nebraska Press.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of applied psychology*, *88*(1), 100-112. https://doi.org/10.1037/0021-9010.88.1.100

Hogan, R., Jones, W. H., & Cheek, J. M. (1985). Socioanalytic theory: An alternative to armadillo psychology. *The Self and Social Life*, *175*, 198.

Höglinger, M., & Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: false positives undermine the crosswise-model RRT. *Political Analysis*, *25*(1), 131-137.

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the randomized response technique and the crosswise model. *Survey Research Methods*, Vol. 10, No. 3, pp. 171-187.

Holbrook, A. L., & Krosnick, J. A. (2010a). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly*, *74*(1), 37-67.

Holbrook, A. L., & Krosnick, J. A. (2010b). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly*, *74*(2), 328-343.

Holden, R. R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *39*(3), 184-201.

Holden, R. R., & Book, A. S. (2012). *Faking does distort self-report personality assessment.* In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (p. 71–84). Oxford University Press.

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology*, *63*(2), 272-279. https://doi.org/10.1037/0022-3514.63.2.272

Holden, R. R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences*, *49*(5), 446–450. https://doi.org/10.1016/j.paid.2010.04.015

Holden, R. R., Wood, L. L., & Tomashewski, L. (2001). Do response time limitations counteract the effect of faking on personality inventory validity?. *Journal of Personality and Social Psychology*, *81*(1), 160-169.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. *ETS Research Report Series*, *1986*(2), i-24.

Holtgraves, T. (2004). Social Desirability and Self-Reports: Testing Models of Socially Desirable Responding. *Personality and Social Psychology Bulletin*, *30*(2), 161–172. https://doi.org/10.1177/0146167203259930

Hooper, A. C. (2007). *Self-presentation on personality measures in lab and field settings: A meta-analysis*. Unpublished doctoral thesis. University of Minnesota.

Hoorens, V., & Buunk, B. P. (1993). Social Comparison of Health Risks: Locus of Control, the Person-Positivity Bias, and Unrealistic Optimism 1. *Journal of Applied Social Psychology*, *23*(4), 291–302.

Hrgović, J., & Hromatko, I. (2019). Self-deception as a function of social status. *Evolutionary Behavioral Sciences*, *13*(3), 223-234.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huang, C. (2013). Relation between self-esteem and socially desirable responding and the role of socially desirable responding in the relation between self-esteem and performance. *European Journal of Psychology of Education*, *28*(3), 663–683. https://doi.org/10.1007/s10212-012-0134-5

Huang, H. Y. (2016). Mixture random-effect IRT models for controlling extreme response style on rating scales. *Frontiers in Psychology*, *7*(NOV), 1–15. https://doi.org/10.3389/fpsyg.2016.01706

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114.

Huber, C. (2017). Faking and the Validity of Personality Tests: Using New Faking-Resistant Measures to Study Some Old Questions. Unpublished doctoral thesis, University of Minnesota.

Hughes, B. L., & Beer, J. S. (2012). Medial orbitofrontal cortex is associated with shifting decision thresholds in self-serving cognition. *NeuroImage*, *61*(4), 889–898. https://doi.org/10.1016/j.neuroimage.2012.03.011

Hull, J. G., & Levy, A. S. (1979). The organizational functions of the self: An alternative to the Duval and Wicklund Model of self-awareness. *Journal of Personality and Social Psychology*, *37*(5), 756.

Hülür, G., Wilhelm, O., & Schipolowski, S. (2011). Prediction of self-reported knowledge with over-claiming, fluid and crystallized intelligence and typical intellectual engagement. *Learning and Individual Differences*, *21*(6), 742–746. https://doi.org/10.1016/j.lindif.2011.09.006

Humberg, S., Dufner, M., Schönbrodt, F. D., Geukes, K., Hutteman, R., Küfner, A. C. P., van Zalk, M. H. W., Denissen, J. J. A., Nestler, S., & Back, M. D. (2019). Is accurate, positive, or inflated self-perception most advantageous for psychological adjustment? A competitive test of key hypotheses. *Journal of Personality and Social Psychology*, *116*(5), 835–859. https://doi.org/10.1037/pspp0000204

Humm, D. G., & Humm, K. A. (1944). Validity of the Humm-Wadsworth Temperament Scale: With Consideration of the Effects of Subjects' Response-Bias. *Journal of Psychology: Interdisciplinary and Applied*, *18*(1), 55–64. https://doi.org/10.1080/00223980.1944.9917207

Humm, D. C., & Humm, K. A. (1947). Compensations for subjects' response-bias in a measure of temperament. *American Psychologist*, *2*, 305.

Hurtz, G. M., & Alliger, G. M. (2002). Influence of coaching on integrity test performance and unlikely virtues scale scores. *Human Performance*, 15, 255–273.

Izdebski, P. (2007). The role of personality and temperamental factors in breast cancer: A 5-year prospective examination. *Polish Psychological Bulletin*, *4*(38), 198-205.

Izdebski, P., Żbikowska, K., & Kotyśko, M. (2013). Przegląd Teorii Aprobaty Społecznej. *Acta Universitatis Lodziensis. Folia Psychologica*, *17*, 5–20. http://hdl.handle.net/11089/4494

Jabkowski, P. (2015). *Reprezentatywność badań reprezentatywnych. Analiza wybranych problemów metodologicznych oraz praktycznych w paradygmacie całkowitego błędu pomiaru*. Poznań: Wydawnictwo Naukowe UAM.

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological bulletin*, *55*(4), 243-252.

Jackson, D. N., & Messick, S. (1962). Response styles and the assessment of psychopathology. *Measurement in personality and cognition*, 129-155.

Jacquemet, N., Joule, R. V., Luchini, S., & Shogren, J. F. (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management*, *65*(1), 110-132.

Jakobsen, M., & Jensen, R. (2015). Common method bias in public management studies. *International Public Management Journal*, *18*(1), 3–30. https://doi.org/10.1080/10967494.2014.997906

Jakubowski, M., Konarzewski, K., Muszyński, M, Smulczyk, M., & Walicki, P. (2017). *Szkolne talenty Europy u progu zmian. Polscy uczniowie w najnowszych badaniach międzynarodowych.* Report on the commission of Fundacja Evidence Institute & Związek Nauczycielstwa Polskiego [Polish Teacher's Union]. Downloaded from: www.evidenceinstitute.pl.

James, W. (1890). *The Principles of Psychology.* Hansie.

Jang, H. (2017). *Answering for Yourself versus Others: Direct versus Indirect Estimates of Charitable Donations*. 1–67.

Jelonek, M., Worek, B., Turek, K. & Muszyński, M. (2019). *Monitoring the lifelong learning in Europe: (in)comparability of indicators and implications for European and local public policies.* Conference presentation: European Survey Research Association (ESRA) 8th Conference, Zagreb, Croatia, July.

Jerrim, J., Luis, Lopez-Agudo, A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). To weight or not to weight?: the case of PISA data. *Investigaciones de Economía de La Educacíon Volume 12*, *Fpu2014 04518*, 285–302.

Jerrim, J., Parker, P., & Shure, N. (2019). Bullshitters. Who Are They and What Do We Know about Their Lives? *Discussion Paper Series*, *April*, 1–34. https://www.iza.org/publications/dp/12282/bullshitters-who-are-they-and-what-do-we-know-about-their-lives

Jo, M. S. (2000). Controlling social-desirability bias via method factors of direct and indirect questioning in structural equation models. *Psychology and Marketing*, *17*(2), 137–148. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2<137::AID-MAR5>3.0.CO;2-V

John, O. P., Cheek, J. M., & Klohnen, E. C. (1996). On the nature of self-monitoring: Construct explication with Q-sort ratings. *Journal of Personality and Social Psychology*, *71*(4), 763.

John, O. P., & Robins, R. W. (1994). Accuracy and bias in self-perception: individual differences in self-enhancement and the role of narcissism. *Journal of personality and social psychology*, *66*(1), 206-219.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of research in personality*, *39*(1), 103-129.

Johnson, D. D. P., & Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, *477*(7364), 317–320. https://doi.org/10.1038/nature10384

Johnson, J. A., & Hogan, R. (2006). A socioanalytic view of faking. [in:] Griffith, R.L. and Mitchell H. Peterson, *A Closer Examination of Applicant Faking Behavior*, 209–231.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36*(2), 264–277. https://doi.org/10.1177/0022022104272905

Johnson, T. P., & Van de Vijver, F. J. (2003). Social desirability in cross-cultural research. *Cross-cultural survey methods*, *325*, 195-204.

Jones, E. E., Gergen, K. J., & Jones, R. G. (1963). Tactics of ingratiation among leaders and subordinates in a status hierarchy. *Psychological Monographs: General and Applied, 77*(3), 1–20. https://doi.org/10.1037/h0093832

Jones, D. N., & Paulhus, D. L. (2009). *Machiavellianism.* In M. R. Leary & R. H. Hoyle (Eds.), *Handbook of individual differences in social behavior* (p. 93–108). The Guilford Press.

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. *Psychological perspectives on the self*, *1*(1), 231-262.

Jones, R. A., Sensenig, J., & Haley, J. V. (1974). Self-descriptions: Configurations of content and order effects. *Journal of Personality and Social Psychology, 30*(1), 36–45. https://doi.org/10.1037/h0036674

Jones, E. E., & Sigall, H. (1971). The bogus pipeline: A new paradigm for measuring affect and attitude. *Psychological Bulletin,* 76(5), 349–364. https://doi.org/10.1037/h0031617

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International.

Joseph, D. L., & Newman, D. A. (2010). Emotional Intelligence: An Integrative Meta-Analysis and Cascading Model. *Journal of Applied Psychology*, *95*(1), 54–78. https://doi.org/10.1037/a0017286

Joseph, J., Berry, K., & Deshpande, S. P. (2009). Impact of emotional intelligence and other factors on perception of ethical behavior of peers. *Journal of Business Ethics*, *89*(4), 539–546. https://doi.org/10.1007/s10551-008-0015-7

Ju, U., & Falk, C. F. (2019). Modeling Response Styles in Cross-Country Self-Reports: An Application of a Multilevel Multidimensional Nominal Response Model. *Journal of Educational Measurement*, *56*(1), 169-191.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*(2), 384–396. https://doi.org/10.1037/0033-295X.107.2.384

Kalton, G., & Schuman, H. (1982). The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society: Series A (General)*, *145*(1), 42–57.

Kam, C. C. S. (2019). Careless responding threatens factorial analytic results and construct validity of personality measure. *Frontiers in Psychology*, *10*(JUN). https://doi.org/10.3389/fpsyg.2019.01258

Kam, C. C. S., & Meyer, J. P. (2015). How Careless Responding and Acquiescence Response Bias Can Influence Construct Dimensionality: The Case of Job Satisfaction. *Organizational Research Methods*, *18*(3), 512–541. https://doi.org/10.1177/1094428115571894

Kam, C., Risavy, S. D., & Perunovic, W. Q. E. (2015). Using over-claiming technique to probe social desirability ratings of personality items: A validity examination. *Personality and Individual Differences*, *74*, 177–181. https://doi.org/10.1016/j.paid.2014.10.017

Kam, C., Schermer, J. A., Harris, J., & Vernon, P. A. (2013). Heritability of acquiescence bias and item keying response style associated with the HEXACO personality scale. *Twin Research and Human Genetics*, *16*(4), 790–798. https://doi.org/10.1017/thg.2013.38

Kaminska, O., & Foulsham, T. (2013). Understanding Sources of Social Desirability Bias in Different Modes : Evidence from Eye-tracking. *Institute for Social & Economic Research*, 1–11.

Kappes, A., & Sharot, T. (2019). The automatic nature of motivated belief updating. *Behavioural Public Policy*, *3*(1), 87–103. https://doi.org/10.1017/bpp.2017.11

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience*, *23*(1), 130–137. https://doi.org/10.1038/s41593-019-0549-2

Kelly, G. A. (1955). *The psychology of personal constructs. Volume 1: A theory of personality*. WW Norton and Company.

Kemmelmeier, M. (2016). Cultural differences in survey responding: Issues and insights in the study of response biases. *International Journal of Psychology*, *51*(6), 439–444. https://doi.org/10.1002/ijop.12386

Kemper, C. J., & Menold, N. (2014). Nuisance or remedy? the utility of stylistic responding as an indicator of data fabrication in surveys. *Methodology*, *10*(3), 92–99. https://doi.org/10.1027/1614-2241/a000078

Kernis, M. H. (2003). Toward a conceptualization of optimal self-esteem. *Psychological Inquiry*, *14*(1), 1–26.

Khorramdel, L. (2014). The influence of different rating scales on impression management in high stakes assessment. *Psychological Test and Assessment Modeling*, *56*(2), 154–167. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc11&NEWS=N&AN=2014-35939-003

Khorramdel, L., & Kubinger, K. D. (2006). The effect of speediness on personality questionnaires: an experiment on applicants within a job recruiting procedure. *Psychology Science*, *48*(3), 378-397.

Khorramdel, L., & von Davier, M. (2014). Measuring Response Styles Across the Big Five: A Multiscale Extension of an Approach Using Multinomial Processing Trees. *Multivariate Behavioral Research*, *49*(2), 161–177. https://doi.org/10.1080/00273171.2013.866536

Khorramdel, L., von Davier, M., & Pokropek, A. (2019). Combining mixture distribution and multidimensional IRTree models for the measurement of extreme response styles. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 538–559. https://doi.org/10.1111/bmsp.12179

Khorramdel, L., von Davier, M., Bertling, J., Roberts, R., & Kyllonen, P. (2017). Recent IRT Approaches to Test and Correct for Response Styles in PISA Background Questionnaire Data: A Feasibility Study. *Psychological Test and Assessment Modeling*, *59*(1), 71.

Kim, Y. H., Chiu, C. Y., & Zou, Z. (2010). Know thyself: misperceptions of actual performance undermine achievement motivation, future performance, and subjective well-being. *Journal of Personality and Social Psychology*, *99*(3), 395–409. https://doi.org/10.1037/a0020555

Klar, Y., & Giladi, E. E. (1997). No one in my group can be below the group's average: A robust positivity bias in favor of anonymous peers. *Journal of personality and social psychology*, *73*(5), 885.

Klar, Y., & Giladi, E. E. (1999). Are most people happier than their peers, or are they just happy?. *Personality and Social Psychology Bulletin*, *25*(5), 586-595.

Klein, S. B., & Lax, M. L. (2010). The unanticipated resilience of trait self-knowledge in the face of neural damage. *Memory*, *18*(8), 918–948.

Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, *17*(2), 161–173.

Klimoski, R., & Hu, X. (2011). Improving self-awareness and self-insight. *Oxford handbook of lifelong learning*, 52-69.

Kofta, M. (1979). *Samokontrola a emocje*. [Self-control and emotions]. Warszawa: Państwowe Wydawnictwo Naukowe (PWN).

Koivula, A., Räsänen, P., & Sarpila, O. (2019, July). Examining Social Desirability Bias in Online and Offline Surveys. In *International Conference on Human-Computer Interaction* (pp. 145-158). Springer, Cham.

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology*, *38*(3), 557-577.

Komar, S., Komar, J. A., Robie, C., & Taggar, S. (2010). Speeding personality measures to reduce faking. Journal of Personnel Psychology, 9, pp. 126-137. https://doi.org/10.1027/1866-5888/a000016

Konarski, R. (2009). *Modele równań strukturalnych: teoria i praktyka.* Warszawa: Wydawnictwo Naukowe PWN.

Konarski, R., & Brycz, H. (2017). Construct and concurrent validity of the Positive Metacognitions and Positive Meta-Emotions Questionnaire in the Polish population. *SAGE Open, 7(2),* 2158244017705423.

Kondratek, B. (2016/2020). "UIRT: Stata module to fit unidimensional Item Response Theory models," Statistical Software Components S458247, Boston College Department of Economics, revised 28 Mar 2020.

Kondratek, B., Skórska, P. & Świst, K. (2015). Wprowadzenie do zróżnicowanego funkcjonowania pozycji testowej. [in:] Artur Pokropek *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii. Teoria i zastosowania.* Warszawa: Instytut Badań Edukacyjnych.

Koniewski, M., Kasparek, K., Czarnik, Sz., & Kocór, M. (2019). *Self-descriptive vs. Objecitve measures of computer skills in the Labour Market Research in Poland.* Conference presentation: European Survey Research Association (ESRA) 8th Conference, Zagreb, Croatia, July.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2), 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Korzeniowski, K. (1980). *Osobowościowe przesłanki spostrzegania osób jako wartości autonomicznej.* [Personality premises on perceiving others as autonomic values]. Unpublished doctoral thesis. Warsaw: Institute of Psychology, University of Warsaw.

Kossowska, M. (2009). Nowe poznawcze wymiary osobowości a społeczne poznanie i działanie. [New cognitive dimensions of personality and social cognition and action.] [in:] Małgorzata Kossowska, Mirosław Kofta (eds.). *Psychologia poznania społecznego: nowe idee*. Warszawa: Wydawnictwa Naukowe PWN.

Kowalski, C. M., Rogoza, R., Vernon, P. A., & Schermer, J. A. (2018). The Dark Triad and the self-presentation variables of socially desirable responding and self-monitoring. *Personality and Individual Differences*, *120*(January), 234–237. https://doi.org/10.1016/j.paid.2017.09.007

Kozielecki, J. (1981). *Psychologiczna teoria samowiedzy.* [Psychological theory of self-knowledge.] Warszawa: PWN.

Krajc, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology*, *29*(5), 724-738.

Kreuter, F. (Ed.). (2013). *Improving surveys with paradata: Analytic uses of process information* (Vol. 581). John Wiley & Sons.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. https://doi.org/10.1093/poq/nfn063

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Krosnick, J. A. (1999). Maximizing questionnaire quality. *Measures of political attitudes*, *2*, 37-58.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly, 52,* 526–538.

Krueger, J. (1998). Enhancement bias in descriptions of self and others. *Personality and Social Psychology Bulletin*, *24*(5), 505–516.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality and Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Krumrei-Mancuso, E. J., Haggard, M. C., LaBouff, J. P., & Rowatt, W. C. (2020). Links between intellectual humility and acquiring knowledge. *Journal of Positive Psychology*, *15*(2), 155–170. https://doi.org/10.1080/17439760.2019.1579359

Krusemark, E. A., Keith Campbell, W., & Clementz, B. A. (2008). Attributions, deception, and event related potentials: an investigation of the self-serving bias. *Psychophysiology*, *45*(4), 511-515.

Kulas, J. T., Klahr, R., & Knights, L. (2019). Confound It! *European Journal of Psychological Assessment*, *35*(6), 855–867. https://doi.org/10.1027/1015-5759/a000459

Kumashiro, M., & Sedikides, C. (2005). Taking on board liability-focused information: Close positive relationships as a self-bolstering resource. *Psychological Science*, *16*(9), 732–739.

Kuncel, N. R., Borneman, M., & Kiger, T. (2012). Innovative item response process and Bayesian faking detection methods: More questions than answers. [in:] Matthias Ziegler, Carolyn MacCann, Richard D. Roberts, *New perspectives on faking in personality assessment*. Oxford University Press.

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology*, *62*(2), 201-228.

Kunda, Z. (1999). *Social cognition: Making sense of people*. MIT press.

Kurman, J. (2001). Self-enhancement: Is it restricted to individualistic cultures?. *Personality and Social Psychology Bulletin*, *27*(12), 1705-1716.

Kurtz, J. E., Tarquini, S. J., & Iobst, E. A. (2008). Socially desirable responding in personality assessment: Still more substance than style. *Personality and Individual Differences*, *45*(1), 22-27.

Kwan, V. S. Y., Barrios, V., Ganis, G., Gorman, J., Lange, C., Kumar, M., Shepard, A., & Keenan, J. P. (2007). Assessing the neural correlates of self-enhancement bias: A transcranial magnetic stimulation study. *Experimental Brain Research*, *182*(3), 379–385. https://doi.org/10.1007/s00221-007-0992-2

Kwan, V. S. Y., John, O. P., Robins, R. W., & Kuang, L. L. (2008). Conceptualizing and assessing self-enhancement bias: A componential approach. *Journal of Personality and Social Psychology*, *94*(6), 1062.

Kwan, V. S. Y., Kuang, L. L., & Hui, N. H. H. (2009). Identifying the sources of self-esteem: The mixed medley of benevolence, merit, and bias. *Self and Identity*, *8*(2–3), 176–195. https://doi.org/10.1080/15298860802504874

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, 277–285. https://doi.org/10.1201/b16061-15

Lamba, S., & Nityananda, V. (2014). Self-deceived individuals are better at deceiving others. *PLoS ONE*, *9*(8), 1–6. https://doi.org/10.1371/journal.pone.0104562

Lancaster, B. P. (1999). *Defining and Interpreting Suppressor Effects: Advantages and Limitations.* Paper presented at the Annual Meeting of the Southwest Educational Research Association (San Antonio, TX, January 21-23, 1999). https://eric.ed.gov/?id=ED426097

Landy, F. J. (1986). Stamp Collecting Versus Science. Validation as Hypothesis Testing. *American Psychologist*, *41*(11), 1183–1192. https://doi.org/10.1037/0003-066X.41.11.1183

Larson, K. E., & Bradshaw, C. P. (2017). Cultural competence and social desirability among practitioners: A systematic review of the literature. *Children and Youth Services Review*, *76*, 100-111.

Leary, M. R. (1990). Responses to social exclusion: Social anxiety, jealousy, loneliness, depression, and low self-esteem. *Journal of Social and Clinical Psychology*, *9*(2), 221-229.

Leary, M.R. (1995). *Self-presentation: Impression Management And Interpersonal Behavior.* Avalon Publishing.

Leary, M. R. (1999). *The social and psychological importance of self-esteem.* In R. M. Kowalski & M. R. Leary (Eds.), *The social psychology of emotional and behavioral problems: Interfaces of social and clinical psychology* (p. 197–221). American Psychological Association. https://doi.org/10.1037/10320-007

Leary, M. R. (2003). Commentary on self-esteem as an interpersonal monitor: The sociometer hypothesis (1995). *Psychological Inquiry*, *14*(3-4), 270-274.

Leary, M. R. (2004). What is the self? A plea for clarity. *Self and Identity*, *3*(1), 1–3.

Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review Psychology*, *58*, 317-344.

Leary, M. R., & Baumeister, R. F. (2000). The nature and function of self-esteem: Sociometer theory. *Advances in Experimental Social Psychology*, *32*, 1–62. https://doi.org/10.1016/s0065-2601(00)80003-9

Leary, M. R., & Tangney, J. P. (2003). The self as an organizing construct in the behavioral and social sciences. *Handbook of Self and Identity*, *15*, 3–14.

Leary, M. R., Tchividijian, L. R., & Kraxberger, B. E. (1994). Self-presentation can be hazardous to your health: Impression management and health risk. *Health Psychology*, *13*(6), 461.

Lee, R. M. (1993). *Doing research on sensitive topics*. Sage.

Lee, C. (2016). *Honesty-Humility and the Overclaiming Technique.* Unpublished master thesis. University of Calgary, Alberta.

Leising, D., Scherbaum, S., Locke, K. D., & Zimmermann, J. (2015). A model of "substance" and "evaluation" in person judgments. *Journal of Research in Personality*, *57*, 61-71.

Leite, W. L., & Beretvas, S. N. (2005). Validation of scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding. *Educational and Psychological Measurement*, *65*(1), 140–154. https://doi.org/10.1177/0013164404267285

Leite, W. L., & Cooper, L. A. (2010). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research*, *45*(2), 271–293. https://doi.org/10.1080/00273171003680245

Lensvelt-Mulders, G. J., Hox, J. J., Van der Heijden, P. G., & Maas, C. J. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research*, *33*(3), 319-348.

Lenzner, T., Kaczmirek, L., & Galesic, M. (2011). Seeing through the eyes of the respondent: An eye-tracking study on survey question comprehension. *International Journal of Public Opinion Research*, *23*(3), 361-373.

Lentz, T. F. (1938). Acquiescence as a factor in the measurement of personality. *Psychological Bulletin*, *35*(9), 659.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255.

Levin, R. A., & Zickar, M. J. (2002). Investigating self-presentation, lies, and bullshit: Understanding faking and its effects on selection decisions using theory, field research, and simulation. *The psychology of work: Theoretically based empirical research*, 253-276.

Levine, E. L., Flory, A., & Ash, R. A. (1977). Self-assessment in personnel selection. *Journal of Applied Psychology*, *62*(4), 428.

Lewicka, M. (1978). Afektywne i deskryptywne formy organizacji informacji w procesie spostrzegania społecznego. Unpublished doctoral thesis, University of Warsaw, Poland.

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment*, *14*(2), 131–141. https://doi.org/10.1111/j.1468-2389.2006.00339.x

Linden, D. E. (2005). The P300: where in the brain is it produced and what does it tell us?. *The Neuroscientist*, *11*(6), 563-576.

Littrell, S., Risko, E.F., & Fugelsang, J.A. (2020). The Bullshitting Frequency Scale: Development and psychometric properties. *British Journal of Social Psychology*, DOI:10.1111/bjso.12379.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88, 767–778.

Loevinger, J., Wessler, R., & Redmore, C. (1970). *Measuring Ego Development. San Fran¬ cisco*. Jossey-Bass.

Logan, D. E., Claar, R. L., & Scharff, L. (2008). Social desirability response bias and self-report of psychological distress in pediatric chronic pain patients. *Pain*, *136*(3), 366-372.

Lönnqvist, J., Verkasalo, M., & Bezmenova, I. (2007). Agentic and communal bias in socially desirable responding. *European Journal of Personality: Published for the European Association of Personality Psychology*, *21*(6), 853–868.

Loosveldt, G., & Beullens, K. (2017). Interviewer effects on non-differentiation and straightlining in the european social survey. *Journal of Official Statistics*, *33*(2), 409–426. https://doi.org/10.1515/jos-2017-0020

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098-2109.

Lorge, I. (1937). Gen-like: Halo or reality. *Psychological Bulletin*, *34*(19371), 545-46.

Lu, Y., & Bolt, D. M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-Scale Assessments in Education*, *3*(1). https://doi.org/10.1186/s40536-015-0012-0

Lucas, R. E., & Baird, B. M. (2004). Extraversion and emotional reactivity. *Journal of personality and social psychology*, *86*(3), 473-485.

Ludeke, S. G., & Makransky, G. (2016). Does the Over-Claiming Questionnaire measure overclaiming? Absent convergent validity in a large community sample. *Psychological Assessment*, *28*(6), 765–774. https://doi.org/10.1037/pas0000211

Ludeke, S. G., Weisberg, Y. J., & Deyoung, C. G. (2013). Idiographically Desirable Responding: Individual Differences in Perceived Trait Desirability Predict Overclaiming. *European Journal of Personality*, *27*(6), 580–592. https://doi.org/10.1002/per.1914

Luo, Y. L. L., Sedikides, C., & Cai, H. (2019). On the Etiology of Self-Enhancement and Its Association With Psychological Well-Being. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550619877410

Lusk, J. L., & Norwood, F. B. (2010). Direct versus indirect questioning: An application to the well-being of farm animals. *Social Indicators Research*, *96*(3), 551–565. https://doi.org/10.1007/s11205-009-9492-z

Łukaszewski, W. (1974). *Osobowość: struktura i funkcje regulacyjne*. [Personality: structure and regulational functions]. Warszawa: Państwowe Wydawnictwo Naukowe (PWN).

Maaß, U., & Ziegler, M. (2017). Narcissistic self-promotion is not moderated by the strength of situational cues. *Personality and Individual Differences*, *104*, 482–488. https://doi.org/10.1016/j.paid.2016.09.008

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, *67*(3), 280–296. https://doi.org/10.1037/0021-9010.67.3.280

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate behavioral research, 38(1),* 113-139.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19-40.

Macdonald, K. (2014). PV: Stata Module to Perform Estimation with Plausible Values, 2014.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in Self-Ratings of Second Language Proficiency: The Role of Language Anxiety. *Language Learning*, *47*(2), 265–287. https://doi.org/10.1111/0023-8333.81997008

Mackinnon, S. P., & Wang, M. (2020). Response-Order Effects for Self-report Questionnaires: Exploring the role of Overclaiming Accuracy and Bias. *Journal of Articles in Support of the Null Hypothesis*, *16*(2).

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.).* Mahwah, New Jersey: Lawrence Erlbaum Associates. [First edition: 1991].

Magdolen, M., Behren, S. Von, Hobusch, J., Chlond, B., & Vortisch, P. (2019). *ScienceDirect Comparison of Response Bias in an Intercultural Context – Evaluation of Psychological Items in Travel Behavior Research*. *00*(May).

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*(3), 252–271. https://doi.org/10.1037/0033-295X.87.3.252

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61-83.

Mansolf, M., & Reise, S. P. (2018). Case diagnostics for factor analysis of ordered categorical data with applications to person-fit measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 86–100.

Maricuțoiu, L. P., & Sârbescu, P. (2019). The relationship between faking and response latencies: A meta-analysis. *European Journal of Psychological Assessment,* 35(1), 3–13. https://doi.org/10.1027/1015-5759/a000361

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*, *84*, 79–83. https://doi.org/10.1016/j.paid.2014.08.021

Marks, G. N., & Pokropek, A. (2019). Family income effects on mathematics achievement: their relative magnitude and causal pathways. *Oxford Review of Education*, *45*(6), 769-785.

Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, *35*(2), 63.

Marselle, T. A. (2014). *A correlational study of explicit and implicit personality tests.* Unpublished doctoral thesis. University of Hartford.

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, 22(2), 366–381. https://doi.org/10.1037/a0019225

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal*, *35*(4), 705–738.

Martial, C., Stawarczyk, D., & D'Argembeau, A. (2018). Neural correlates of context-independent and context-dependent self-knowledge. *Brain and Cognition*, *125*(October 2017), 23–31. https://doi.org/10.1016/j.bandc.2018.05.004

Martocchio, J. J., & Judge, T. A. (1997). Relationship between conscientiousness and learning in employee training: Mediating influences of self-deception and self-efficacy. *Journal of Applied Psychology*, *82*(5), 764.

Maydeu-Olivares, A. (2005). Further empirical results on parametric versus non-parametric IRT modeling of likert-type personality data. *Multivariate Behavioral Research*, *40*(2), 261–279. https://doi.org/10.1207/s15327906mbr4002_5

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

Maydeu-Olivares, A. (2015). Evaluating fit in IRT models. [in:] Steve P . Reise & Dennis A . Revicki (Eds.). *Handbook of Item Response Theory Modeling : Applications to Typical Performance Assessment* (pp . 111-127). Routledge.

Mazza, C., Monaro, M., Burla, F., Colasanti, M., Orrù, G., Ferracuti, S., & Roma, P. (2020). Use of mouse-tracking software to detect faking-good behavior on personality questionnaires: an explorative study. *Scientific Reports*, *10*(1), 1–13. https://doi.org/10.1038/s41598-020-61636-5

McAdams, DP (1985). The "imago": A key narrative component of identity. In *Review of personality and social psychology* (Vol. 6, pp. 114-141). Beverly Hills, CA: Sage.

McClelland, G. H., Lynch, J. G., Irwin, J. R., Spiller, S. A., & Fitzsimons, G. J. (2015). Median splits, Type II errors, and false-positive consumer psychology: Don't fight the power. *Journal of Consumer Psychology*, *25*(4), 679–689. https://doi.org/10.1016/j.jcps.2015.05.006

McCrae, R. R. (2018). Method biases in single-source personality assessments. *Psychological Assessment*, *30*(9), 1160–1173. https://doi.org/10.1037/pas0000566

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, *51*(6), 882–888. https://doi.org/10.1037/0022-006X.51.6.882

McCutcheon, A. L. (1987). *Latent class analysis (No. 64).* Newbury Park: Sage Publications.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85*(5), 812–821. https://doi.org/10.1037/0021-9010.85.5.812

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology*, *36*(4), 979–1016. https://doi.org/10.1111/j.0021-9029.2006.00052.x

McGuire, W. J. (1968). Personality and attitude change: An information-processing theory. *Psychological foundations of attitudes*, *171*, 196.

McIntosh, R. D., Fowler, E. A., Lyu, T., & Sala, S. Della. (2019). Wise Up: Clarifying the Role of Metacognition in the Dunning-Kruger Effect. *Journal of Experimental Psychology: General*, 1–43. https://doi.org/10.1037/xge0000579

McLarres, P., & Oyelere, P. (1999). A critical analysis of self-assessed entry-level personal computer skills among newly-qualified Irish chartered accountants. *Accounting Education*, *8*(3), 203–216.

McNicol, D. (1972/2005). *A primer of signal detection theory*. London: George Allen & Unwin.

Mead, G. H. (1934). *Mind, Self & Society from the Standpoint of a Social Behaviorist*. Chicago, Ill: The University of Chicago Press.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meade, A. W., Pappalardo, G., Braddy, P. W., & Fleenor, J. W. (2020). Rapid Response Measurement: Development of a Faking-Resistant Assessment Method for Personality. *Organizational Research Methods*, *23*(1), 181-207.

Meehl, P. E., & Hathaway, S. R. (1946). The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology, 30*(5), 525–564. https://doi.org/10.1037/h0053634

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education?. *Personality and individual differences*, *44*(7), 1539-1550.

Melchers, M., Plieger, T., Montag, C., Reuter, M., Spinath, F. M., & Hahn, E. (2018). The heritability of response styles and its impact on heritability estimates of personality: A twin study. *Personality and Individual Differences*, *134*(February), 16–24. https://doi.org/10.1016/j.paid.2018.05.023

Mele, A. R. (1997). Real self-deception. *Behavioral and Brain Sciences*, *20*(1), 91–102.

Menold, N., & Kemper, C. J. (2014). How do real and falsified data differ? psychology of survey response as a source of falsification indicators in facel-to-Face surveys. *International Journal of Public Opinion Research*, *26*(1), 41–65. https://doi.org/10.1093/ijpor/edt017

Merlo, J., Wagner, P., Austin, P. C., Subramanian, S. V., & Leckie, G. (2018). General and specific contextual effects in multilevel regression analyses and their paradoxical relationship: A conceptual tutorial. *SSM - Population Health*, *5*(March), 33–37. https://doi.org/10.1016/j.ssmph.2018.05.006

Mesmer-Magnus, J., Viswesvaran, C., Deshpande, S., & Joseph, J. (2006). Social desirability: The role of over-claiming, self-esteem, and emotional intelligence. *Psychology Science*, *48*(3), 336-356.

Messick, S. (1960). Dimensions of social desirability. *Journal of Consulting Psychology*, *24*(4), 279–287. https://doi.org/10.1037/h0044153

Messick, S. (1987). Validity. *ETS Research Report Series*, *1987*(2), i-208.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227–234. https://doi.org/10.1037/h0031564

Mijović-Prelec, D., & Prelec, D. (2010). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1538), 227–240.

Milgram, S. (1961). Nationality and conformity. *Scientific American*, *205*(6), 45-51.

Miller (Droitcour), J. D. (1985). The nominative technique: A new method of estimating heroin prevalence. *NIDA Research Monograph*, *54*, 104-124.

Miller, C. E., & Barrett, G. V. (2001). *The coachability and fakability of personality selection tests used for police selection.* Paper presented at the 25th annual conference of the International Personnel Management Association Assessment Council, Newport Beach, CA.

Miller, T. M., & Geraci, L. (2011). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning*, *6*(3), 303–314. https://doi.org/10.1007/s11409-011-9083-7

Miller, D. T., & Ratner, R. K. (1998). The disparity between the actual and assumed power of self-interest. *Journal of personality and social psychology*, *74*(1), 53-62.

Miller, N., Resnick, P., & Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, *51*(9), 1359-1373.

Mills, C., & Hogan, R. (1978). A role theoretical interpretation of personality scale item responses. *Journal of Personality*, *46*(4), 778–785. https://doi.org/10.1111/j.1467-6494.1978.tb00197.x

Minkov, M. (2008). Self-enhancement and self-stability predict school achievement at the national level. *Cross-Cultural Research*, *42*(2), 172–196. https://doi.org/10.1177/1069397107312956

Mitchell, T. W. & Klimoski, R. J. (1982). Is it Rational to be Empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology*, 67(4), 411–418

Moehring, K., & Schmidt-Catran, A. (2013). MLT: Stata module to provide multilevel tools. https://econpapers.repec.org/software/bocbocode/s457577.htm

Moore, D. A. (2005). Myopic biases in strategic social prediction: Why deadlines put everyone under more pressure than everyone else. *Personality and Social Psychology Bulletin*, *31*(5), 668-679.

Moore, D. A., & Cain, D. M. (2004). *Myopic biases in comparative judgment and entrepreneurial entry*. Tepper Working Paper.

Moore, D. A., & Dev, A. S. (2017). Individual differences in overconfidence. *Encyclopedia of Personality and Individual Differences. Springer. Retrieved from Http://Osf. Io/Hzk6q*.

Moore, D. A., & Healy, P. J. (2008). The Trouble With Overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology*, *65*(2), 131–149. https://doi.org/10.1111/j.2044-8325.1992.tb00490.x

Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: a meta-analysis. *Psychological Bulletin*, *128*(4), 638.

Morey, L. C., & Lanier, V. W. (1998). Operating characteristics of six response distortion indicators for the Personality Assessment Inventory. *Assessment*, *5*(3), 203–214.

Morren, M., Gelissen, J., & Vermunt, J. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology*, 8, pp. 159-170. https://doi.org/10.1027/1614-2241/a000048

Mõttus, R., Allik, J., & Realo, A. (2020). Do self-reports and informant-ratings measure the same personality constructs? *European Journal of Psychological Assessment, 36*(2), 289–295. https://doi.org/10.1027/1015-5759/a000516

Moutafi, J., Furnham, A., & Crump, J. (2006). What facets of openness and conscientiousness predict fluid intelligence score?. *Learning and Individual Differences*, *16*(1), 31-42.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, George C., I. (2006). Individual differences in impression management: an exploration of the psychological processes underlying faking. *Psychology Science*, *48*(3), 288–312. http://www.pabst-publishers.de/psychology-science/3-2006/ps_3_2006_288-312.pdf

Muller, S. (2019). *How to (Not) Measure Self-Favoring Response Bias – An Examination of Impression Management and Overclaiming*. Unpublished doctoral dissertation, Ulm University.

Müller, S., & Moshagen, M. (2018). Overclaiming shares processes with the hindsight bias. *Personality and Individual Differences*, *134*(June), 298–300. https://doi.org/10.1016/j.paid.2018.06.035

Müller, S., & Moshagen, M. (2019a). Controlling for Response Bias in Self-Ratings of Personality: A Comparison of Impression Management Scales and the Overclaiming Technique. *Journal of Personality Assessment*, *101*(3), 229–236. https://doi.org/10.1080/00223891.2018.1451870

Müller, S., & Moshagen, M. (2019b). True virtue, self-presentation, or both?: A behavioral test of impression management and overclaiming. *Psychological assessment*, *31*(2), 181-191.

Müller, F., Schiepe-Tiska, A., & Strohmaier, A. R. (2017). EyeQuest–Cross-cultural comparison of eye movements and self-reports. *Vortrag auf dem jährlichen Treffen des National Council on Measurement in Education (NCME), San Antonio, TX, USA. Zugriff am*, *4*, 2019.

Murphy, G. (1947). *Personality: A biosocial approach to origins and structure.* Harper & Brothers. https://doi.org/10.1037/10759-000

Murray, C. J. L., Ozaltin, E., Tandon, A., Salomon, J., Sadana, R., & Chatterji, S. (2003). Empirical evaluation of the anchoring vignette approach in health surveys. *Health Systems Performance Assessment: Debates, Methods and Empiricism*, *369*, 399.

Musch, J. (2003). Personality differences in hindsight bias. *Memory*, *11*(4-5), 473-489.

Musch, J., Ostapczuk, M., & Klaiber, Y. (2012). Validating an inventory for the assessment of egoistic bias and moralistic bias as two separable components of social desirability. *Journal of Personality Assessment*, *94*(6), 620–629. https://doi.org/10.1080/00223891.2012.672505

Muszyński, M., Campfield, D.E., & Szpotowicz, M. (2015). *Język angielski w szkole podstawowej–proces i efekty nauczania*. Warszawa: Instytut Badań Edukacyjnych.

Muthén, B.O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52(3), 431-462.

Muthén, L.K. and Muthén, B.O. (1998-2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA: Muthén & Muthén.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*(1), 261–286. https://doi.org/10.1111/j.1467-6494.2008.00545.x

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280. https://doi.org/10.1002/ejsp.2420150303

Nias, D. K. B. (1972). The effects of providing a warning about the lie scale in a personality inventory. *British Journal of Educational Psychology*, *42*(3), 308-312.

Nichols, R. M., & Loftus, E. F. (2019). Who is susceptible in three false memory tasks? *Memory*, *27*(7), 962–984. https://doi.org/10.1080/09658211.2019.1611862

Niedźwieńska, A. (2003). Gender differences in vivid memories. *Sex Roles*, *49*(7-8), 321-331.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2017). Measuring non-cognitive predictors in high-stakes contexts: The effect of self-presentation on self-report instruments used in admission to higher education. *Personality and Individual Differences*, *106*, 183–189. https://doi.org/10.1016/j.paid.2016.11.014

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of communication*, *24*(2), 43-51.

Noelle-Neumann, E. (1991). The Theory of Public Opinion: The Concept of the Spiral of Silence. *Annals of the International Communication Association*, *14*(1), 256–287. https://doi.org/10.1080/23808985.1991.11678790

Nuhfer, E., Cogan, C., Fleischer, S., Gaze, E., & Wirth, K. (2016). Random Number Simulations Reveal How Random Noise Affects the Measurements and Graphical Portrayals of Self-Assessed Competency. *Numeracy*, *9*(1). https://doi.org/10.5038/1936-4660.9.1.4

Nuhfer, E., Fleischer, S., Cogan, C., Wirth, K., & Gaze, E. (2017). How Random Noise and a Graphical Convention Subverted Behavioral Scientists' Explanations of Self-Assessment Data: Numeracy Underlies Better Alternatives. *Numeracy*, *10*(1). https://doi.org/10.5038/1936-4660.10.1.4

Nygren, T., & Guath, M. (2019). Swedish teenagers' difficulties and abilities to determine digital news credibility. *Nordicom Review*, *40*(1), 23–42. https://doi.org/10.2478/nor-2019-0002.

O'Dell, J. W. (1971). Method for detecting random answers on personality questionnaires. *Journal of Applied Psychology*, *55*(4), 380–383. https://doi.org/10.1037/h0031473

OECD (2014a), *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science* (Volume I, Revised edition, February 2014), PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264201118-en

OECD (2014b). *PISA 2012 Technical Report*. PISA, OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD (2019), How much effort did students invest in the PISA test?, [in] *PISA 2018 Results (Volume I): What Students Know and Can Do*, OECD Publishing, Paris. https://doi.org/10.1787/04fd5153-en

Oh, I.-S., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: a meta-analysis. *Journal of Applied Psychology*, *96*(4), 762–773. https://doi.org/10.1037/a0021832

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, *11*(2–3), 245–269.

Ones, D. S., & Viswesvaran, C. (1999). Relative importance of personality dimensions for expatriate selection: A policy capturing study. *Human performance*, *12*(3-4), 275-294.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*(6), 660–679.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology*, *1*, 220.

Pace, V. L., & Borman, W.C. (2006). The use of warnings to discourage faking on non-cognitive inventories. [in:] Richard L. Griffith & Mitchell M. Peterson, *A closer examination of applicant faking behavior*, Greenwich, Connecticut, IAP, 21-42.

Pannone, R. D. (1984). Predicting Test Performance: a Content Valid Approach To Screening Applicants. *Personnel Psychology*, *37*(3), 507–514. https://doi.org/10.1111/j.1744-6570.1984.tb00526.x

Papps, B. P., & O'Carroll, R. E. (1998). Extremes of self-esteem and narcissism and the experience and expression of anger and aggression. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *24*(6), 421–438.

Paulewicz, B., & Blaut, A. (2020). The bhsdtr package: a general-purpose method of Bayesian inference for Signal Detection Theory models. *Behavior Research Methods*, 1-20.

Paulhus, D. L. (1981). Control of social desirability in personality inventories: Principal-factor deletion. *Journal of Research in Personality*, *15*(3), 383–388. https://doi.org/10.1016/0092-6566(81)90035-0

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L. (1986). Self-Deception and Impression Management in Test Responses. *Personality Assessment via Questionnaires*, 143–165. https://doi.org/10.1007/978-3-642-70751-3_8

Paulhus, D. (1991). Measurement and control of response bias. [in:] J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press, Inc.

Paulhus, D. L. (2002). Socially Desirable Responding: The Evolution of a Construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (Issue 2002, pp. 49–69). Erlbaum. https://doi.org/10.1097/00005053-195311720-00010

Paulhus, D. L. (2011). Overclaiming on personality questionnaires. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 151-164). New York, NY: Oxford University Press

Paulhus, D. (2017). Encyclopedia of Personality and Individual Differences. *Encyclopedia of Personality and Individual Differences*, *January*. https://doi.org/10.1007/978-3-319-28099-8

Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin*, *21*(2), 100-108.

Paulhus, D. L., & Dubois, P. J. (2014). Application of the Overclaiming Technique to Scholastic Assessment. *Educational and Psychological Measurement*, *74*(6), 975–990. https://doi.org/10.1177/0013164414536184

Paulhus, D. L., Graf, P., & Van Selst, M. (1989). Attentional Load Increases the Positivity of Self-Presentation. *Social Cognition*, *7*(4), 389–400. https://doi.org/10.1521/soco.1989.7.4.389

Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence*, *32*(3), 297–314. https://doi.org/10.1016/j.intell.2004.02.001

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The Over-Claiming Technique: Measuring Self-Enhancement Independent of Ability. *Journal of Personality and Social Psychology*, *84*(4), 890–904. https://doi.org/10.1037/0022-3514.84.4.890

Paulhus, D. L., & John, O. P. (1998). Egoistic and Moralistic Biases in Self-Perception: The Interplay of Self-Deceptive Styles with Basic Traits and Motives. *Journal of Personality*, *66*(6), 1025–1060. https://doi.org/10.1111/1467-6494.00041

Paulhus, D. L., & Levitt, K. (1987). Desirable Responding Triggered by Affect: Automatic Egotism? *Journal of Personality and Social Psychology*, *52*(2), 245–259. https://doi.org/10.1037/0022-3514.52.2.245

Paulhus, D. L., Lysy, D. C., & Yik, M. S. M. (1998). Self-Report Measures of Intelligence: Are They Useful as Proxy IQ Tests? *Journal of Personality*, *66*(4), 525–554. https://doi.org/10.1111/1467-6494.00023

Paulhus, D. L., & Notareschi, R. F. (1993). Varieties of faking manipulations. *Unpublished data, University of British Columbia*.

Paulhus, D. L., & Petrusic, W. M. (2010). Measuring individual differences with signal detection analysis: A guide to indices based on knowledge ratings. *Manuscript under review*.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and Denial in Socially Desirable Responding. *Journal of Personality and Social Psychology*, *60*(2), 307–317. https://doi.org/10.1037/0022-3514.60.2.307

Paulhus, D. L., Robins, R. W., Trzesniewski, K. H., & Tracy, J. L. (2004). Two replicable suppressor situations in personality research. *Multivariate Behavioral Research*, *39*(2), 303-328.

Paulhus, D. L., & Trapnell, P. D. (2008). *Self-presentation of personality: An agency-communion framework.* In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (p. 492–517). The Guilford Press.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. [in:] Richard W. Robins, R. Chris Fraley, Robert F. Krueger, *Handbook of research methods in personality psychology*, *1*, 224-239.

Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*, *36*(6), 556–563. https://doi.org/10.1016/S0092-6566(02)00505-6

Pauls, C. A., & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences*, *37*(6), 1137–1151. https://doi.org/10.1016/j.paid.2003.11.018

Pauls, C. A., & Stemmler, G. (2003). Substance and bias in social desirability responding. *Personality and Individual Differences*, *35*(2), 263–275. https://doi.org/10.1016/S0191-8869(02)00187-3

Paunonen, S. V. (2016). Sex Differences in Judgments of Social Desirability. *Journal of Personality*, *84*(4), 423–432. https://doi.org/10.1111/jopy.12169

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). Effects of Applicant Faking on Forced-Choice and Likert Scores. In *Organizational Research Methods* (Vol. 22, Issue 3). https://doi.org/10.1177/1094428117753683

Pedregon, C. A., Farley, R. L., Davis, A., Wood, J. M., & Clark, R. D. (2012). Social desirability, personality questionnaires, and the " better than average" effect. *Personality and Individual Differences*, *52*(2), 213–217. https://doi.org/10.1016/j.paid.2011.10.022

Pelt, D. H. M., Van der Linden, D., Dunkel, C. S., & Born, M. P. (2019). The Motivation and Opportunity for Socially Desirable Responding Does Not Alter the General Factor of Personality. *Assessment*. https://doi.org/10.1177/1073191119880960

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, *10*(6), 549–563. https://doi.org/10.5281/zenodo.1067051

Perinelli, E., & Gremigni, P. (2016). Use of Social Desirability Scales in Clinical Psychology: A Systematic Review. *Journal of Clinical Psychology*, *72*(6), 534–551. https://doi.org/10.1002/jclp.22284

Pesta, B. J., & Poznanski, P. J. (2009). The inspection time and over-claiming tasks as predictors of MBA student performance. *Personality and Individual Differences*, *46*(2), 236–240. https://doi.org/10.1016/j.paid.2008.10.005

Peterson, J. B. (1999). *Maps of meaning: The architecture of belief*. New York: Routledge.

Peterson, J. B., DeYoung, C. G., Driver-Linn, E., Séguin, J. R., Higgins, D. M., Arseneault, L., & Tremblay, R. E. (2003). Self-deceptive and failure to modulate responses despite accruing evidence of error. *Journal of Research in Personality*, *37*(3), 205–223. https://doi.org/10.1016/S0092-6566(02)00569-X

Peterson, J. B., Driver-Linn, E., & DeYoung, C. G. (2002). Self-deception and impaired categorization of anomaly. *Personality and Individual Differences*, *33*(2), 327–340. https://doi.org/10.1016/S0191-8869(01)00158-1

Peterson, R. A., Rhi-Perez, P., & Albaum, G. (2014). A cross-national comparison of extreme response style measures. *International Journal of Market Research*, *56*(1), 89–110. https://doi.org/10.2501/IJMR-2014-005

Petrocelli, J. V. (2018). Antecedents of bullshitting. *Journal of Experimental Social Psychology*, *76*(March), 249–258. https://doi.org/10.1016/j.jesp.2018.03.004

Phares, J. E., Ritchie, E. D., & Davis, W. L. (1968). Internal-external control and reaction to threat. *Journal of Personality and Social Psychology*, *10*(4), 402.

Phillips, D. L., & Clancy, K. J. (1972). Some Effects of "Social Desirability" in Survey Studies. *American Journal of Sociology*, *77*(5), 921–940. https://doi.org/10.1086/225231

Piedmont, R. L., McCrae, R. R., Rieman, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, *78*(3), 582–593. https://doi.org/10.1037/0022-3514.78.3.582

Plieninger, H. (2020a). Developing and applying IR-tree models: Guidelines, caveats, and an extension to multiple groups. *Organizational Research Methods*, 1094428120911096.

Plieninger, H. (2020b). R package 'ItemResponseTrees'. https://cran.r-project.org/web/packages/ItemResponseTrees/ItemResponseTrees.pdf

Plieninger, H., & Meiser, T. (2014). Validity of Multiprocess IRT Models for Separating Content and Response Styles. *Educational and Psychological Measurement*, *74*(5), 875–899. https://doi.org/10.1177/0013164413514998

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Pokropek, A. (2011). Missing by Design: Planned Missing-Data Designs in Social Science. *Institute of Philosophy and Sociology Polish Academy of Sciences*, *20*(1), 81–105. www.ifispan.waw.pl

Pokropek, A. (2014). *Dekonstrukcja skal szacunkowych . Przykład skali znajomości pojęć matematycznych uczniów w PISA 2012* (pp. 119–127). http://www.ptde.org/pluginfile.php/879/mod_page/content/2/Archiwum/XX_KDE/pdf_2014/Pokropek.pdf

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics*, 41(3), 300-325.

Pokropek, A., Khorramdel, L., & von Davier, M. (in preparation). Evaluation of an IRTree Approach to Detect Response Styles: Simulation Study and Empirical Evidence. *Unpublished manuscript.*

Polczyk, R. (2005). Interrogative suggestibility: Cross-cultural stability of psychometric and correlational properties of the Gudjonsson Suggestibility Scales. *Personality and Individual Differences*, *38*(1), 177-186.

Pozzoli, T., Gini, G., & Vieno, A. (2012). The role of individual correlates and class norms in defending and passive bystanding behavior in bullying: A multilevel analysis. *Child Development*, *83*(6), 1917–1931.

Prelec, D. (2004). A Bayesian truth Serum for subjective data. *Science*, *306*(5695), 462–466. https://doi.org/10.1126/science.1102081

Pyszczynski, T., Greenberg, J., Solomon, S., Arndt, J., & Schimel, J. (2004). Why do people need self-esteem? A theoretical and empirical review. *Psychological Bulletin*, *130*(3), 435.

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology*, *46*(2), 237.

Raghavarao, D., & Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(1), 40-45.

Rahnev, D., Desender, K., Lee, A. L. F., Adler, W. T., Aguilar-Lleyda, D., Akdoğan, B., Arbuzova, P., Atlas, L. Y., Balcı, F., Bang, J. W., Bègue, I., Birney, D. P., Brady, T. F., Calder-Travis, J., Chetverikov, A., Clark, T. K., Davranche, K., Denison, R. N., Dildine, T. C., … Zylberberg, A. (2020). The Confidence Database. *Nature Human Behaviour*, *4*(3), 317–325. https://doi.org/10.1038/s41562-019-0813-1

Randall, D.M., Fernandes, M.F. (1991). The social desirability response bias in ethics research. *J Bus Ethics 10,* 805–817. https://doi.org/10.1007/BF00383696

Rasinski, K. A., Baldwin, A. K., Willis, G. B., & Jobe, J. B. (1994). Risk and loss perceptions associated with survey reporting of sensitive behaviors. In *annual meeting of the American Statistical Association, Toronto, Canada*.

Rasinski, K. A., Visser, P. S., Zagatsky, M., & Rickett, E. M. (2005). Using implicit goal priming to improve the quality of self-report data. *Journal of Experimental Social Psychology*, *41*(3), 321-327.

Raskin, R., & Hall, C. S. (1981). The Narcissistic Personality Inventory: Alternative form reliability and further evidence of construct validity. *Journal of personality assessment*, *45*(2), 159-162.

Rauch, W. A., Schweizer, K., & Moosbrugger, H. (2007). Method effects due to social desirability as a parsimonious explanation of the deviation from unidimensionality in LOT-R scores. *Personality and Individual Differences*, *42*(8), 1597-1607.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Newbury Park, CA: Sage.

Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of educational and behavioral statistics*, *20*(4), 307-335.

Rauthmann, J. F. (2011). Acquisitive or protective self-presentation of dark personalities? Associations among the Dark Triad and self-monitoring. *Personality and Individual Differences*, *51*(4), 502–508.

Reeder, M., C., & Ryan, M. (2012). Methods for correcting for faking. [in:] Matthias Ziegler, Carolyn MacCann, Richard D. Roberts (Eds.), *New perspectives on faking in personality assessment,* Oxford University Press, 131-151.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment,* 95(2), 129-140.

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the Bifactor Model a Better Model or Is It Just Better at Modeling Implausible Responses? Application of Iteratively Reweighted Least Squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, *51*(6), 818–838. https://doi.org/10.1080/00273171.2016.1243461

Revelle, W., & Wilt, J. (2013). The general factor of personality: A general critique. *Journal of research in personality*, *47*(5), 493-504.

Reynolds, C. R., & Suzuki, L. A. (2013). Bias in psychological assessment: An empirical review and recommendations. In J. R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology: Assessment psychology*., Vol. 10, 2nd ed. (pp. 82–113). Hoboken, NJ: John Wiley & Sons Inc.

Rhodewalt, F., & Morf, C. C. (1995). Self and interpersonal correlates of the Narcissistic Personality Inventory: A review and new findings. Journal of Research in Personality, 29(1), 1–23. https://doi.org/10.1006/jrpe.1995.1001

Richman, W. L., Weisband, S., Kiesler, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*(5), 754–775. https://doi.org/10.1037/0021-9010.84.5.754

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*(5695), 443–447.

Rigney, A. E. (2019). *The role of biased searching through memory in motivated social evaluation*. Unpublished doctoral dissertation. University of Texas, TX, USA.

Riketta, M. (2004). Does social desirability inflate the correlation between self-esteem and anxiety? *Psychological Reports*, *94*(3_suppl), 1232–1234.

Robie, C., Komar, S., & Brown, D. J. (2010). The effects of coaching and speeding on big five and impression management scale scores. *Human Performance*, *23*(5), 446–467. https://doi.org/10.1080/08959285.2010.515278

Robie, C., Taggar, S., & Brown, D. J. (2009). The effects of warnings and speeding on scale scores and convergent validity of conscientiousness. *Human Performance*, *22*(4), 340-354.

Robins, R. W., & Beer, J. S. (2001). Positive illusions about the self: Short-term benefits and long-term costs. *Journal of Personality and Social Psychology*, *80*(2), 340–352. https://doi.org/10.1037/0022-3514.80.2.340

Robins, R. W., & Paulhus, D. L. (2004). The character of self-enhancers: Implications for organizations. *Personality Psychology in the Workplace.*, 193–219. https://doi.org/10.1037/10434-008

Robins, R. W., Tracy, J. L., Trzesniewski, K., Potter, J., & Gosling, S. D. (2001). Personality correlates of self-esteem. *Journal of Research in Personality*, *35*(4), 463–482. https://doi.org/10.1006/jrpe.2001.2324

Robins, R., & John, O. (1997). The quest for self-insight: Theory and research on accuracy and bias in self-perception. *Handbook of Personality Psychology*, 649–679. http://psycnet.apa.org/psycinfo/1997-08808-025

Robinson, D. L. (2001). How brain arousal systems determine different temperament types and the major dimensions of personality. *Personality and Individual Differences*, 31(8), 1233-1259.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21(2), 137–150. https://doi.org/10.1037/met0000045

Roediger III, H. L. (1996). Memory illusions. *Journal of memory and Language*, *35*(2), 76-100.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of experimental psychology: Learning, Memory, and Cognition*, *21*(4), 803-814.

Roediger, H. L., & McDermott, K. B. (2000). Distortions of memory. *The Oxford handbook of memory*, 149-162.

Rogers, R. & Bender, S.D. (Eds.), *Clinical assessment of malingering and deception.* New York, NY, US: The Guilford Press.

Rogers, R., Gillis, J. R., Bagby, R. M., & Monteiro, E. (1991). Detection of malingering on the Structured Interview of Reported Symptoms (SIRS): A study of coached and uncoached simulators. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*(4), 673–677. https://doi.org/10.1037/1040-3590.3.4.673

Roma, P., Verrocchio, M. C., Mazza, C., Marchetti, D., Burla, F., Cinti, M. E., & Ferracuti, S. (2018). Could time detect a faking-good attitude? A study with the MMPI-2-RF. *Frontiers in Psychology*, *9*(JUL), 1–9. https://doi.org/10.3389/fpsyg.2018.01064

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin, 63*(3), 129–156. https://doi.org/10.1037/h0021888

Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). *Acceptance and commitment therapy. Measures package*, *61*(52), 18.

Rosenthal, S. A., Hooley, J. M., & Steshenko, Y. (2007). Distinguishing grandiosity from self-esteem: Development of the Narcissistic Grandiosity Scale. *Manuscript in preparation*.

Ross, S. R., Canada, K. E., & Rausch, M. K. (2002). Self-handicapping and the five factor model of personality: Mediation between neuroticism and conscientiousness. *Personality and Individual Differences*, *32*(7), 1173–1184.

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology*, *13*(3), 279-301.

Ross, C. E., & Mirowsky, J. (1984). Socially-desirable response and acquiescence in a cross-cultural survey of mental health. *Journal of Health and Social Behavior*, 189-197.

Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*, *6*(3), 376–400. https://doi.org/10.1093/JSSAM/SMX020

Roth, D. L., Snyder, C. R., & Pace, L. M. (1986). Dimensions of favorable self-presentation. *Journal of Personality and Social Psychology*, *51*(4), 867.

Roulin, N., & Krings, F. (2016). When winning is everything: The relationship between competitive worldviews and job applicant faking. *Applied Psychology*, *65*(4), 643–670.

Różycka-Tran, J., & Wojciszke, B. (2010). Skala wiary w grę o sumie zerowej. *Studia Psychologiczne*, *48*(4), 35-46.

Rubin, D. B. (1996). Multiple Imputation after 18+ Years. In *Journal of the American Statistical Association* (Vol. 91, Issue 434, pp. 473–489). https://doi.org/10.1080/01621459.1996.10476908

Rubinstein, S.L. (1960). *Problems of general psychology.* Moscow: Pedagogika.

Ruch, F. L. (1942). A technique for detecting attempts to fake performance on a self-inventory type of personality test. *Studies in personality*, 229-234.

Rudman, L. A. (1998). Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology*, *74*(3), 629.

Rundquist, E. A., & Sletto, R. F. (1936). *Personality in the Depression: A Study in the Measurement of Attitudes* (No. 12). Minneapolis: University of Minnesota Press.

Rutkowski, D., & Wild, J. (2015). Stakes matter: Student motivation and the validity of student assessments for teacher evaluation. *Educational Assessment*, *20*(3), 165–179.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Ryals, A. J., Yadon, C. A., Nomi, J. S., & Cleary, A. M. (2011). When word identification fails: ERP correlates of recognition without identification and of word identification failure. *Neuropsychologia*, *49*(12), 3224–3237. https://doi.org/10.1016/j.neuropsychologia.2011.07.027

Rynko, M. & Palczyńska, M. (2018). The quality of ICT skills indicators. *Unpublished conference paper: European Conference in Official Statistics, Krakow, Poland, June.*

Ryvkin, D., Krajč, M., & Ortmann, A. (2012). Are the unskilled doomed to remain unaware? *Journal of Economic Psychology*, *33*(5), 1012–1031.

Sackeim, H. A., & Gur, R. C. (1978). Self-deception, self-confrontation, and consciousness. [in:] G.E. Schwartz and D. Shapiro (Eds.), *Consciousness and self-regulation* (pp. 139-197). Springer, Boston, MA.

Sackeim, H. A., & Gur, R. C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology, 47*(1), 213–215. https://doi.org/10.1037/0022-006X.47.1.213

Saphire-Bernstein, S., Way, B. M., Kim, H. S., Sherman, D. K., & Taylor, S. E. (2011). Oxytocin receptor gene (OXTR) is related to psychological resources. *Proceedings of the National Academy of Sciences*, *108*(37), 15118-15122.

Sârbescu, P., Costea, I., & Rusu, S. (2012). Psychometric properties of the Marlowe-Crowne Social Desirability Scale in a Romanian sample. *Procedia - Social and Behavioral Sciences*, *33*, 707–711. https://doi.org/10.1016/j.sbspro.2012.01.213

Sassenrath, C. (2019). “Let Me Show You How Nice I Am”: Impression Management as Bias in Empathic Responses. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550619884566

Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management journal*, *43*(6), 1248-1264.

Schaeffer, N. C., & Charng, H.-W. (1991). Two experiments in simplifying response categories: intensity ratings and behavioral frequencies. *Sociological Perspectives*, *34*(2), 165–182.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online, 8(2),* 23-74.

Schilling, M., Sparfeldt, J. R., Becker, N., Engel, M., Levacher, J., Tilman, S. F. P., Schafer, J., Schwabe, S., & Koenig, C. J. (2020). Is it enough to be willing to win or do you have to be smart ? The relationship between competitive worldviews , cognitive abilities , and applicant faking in personality tests. *International Journal of Selection and Assessement*, *In PRESS*(May).

Schlenker, B. R. (1980). *Impression management*. Monterey, CA: Brooks/Cole Publishing Company.

Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal processes involving impression regulation and management. *Annual Review of Psychology*, *43*(1), 133–168.

Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the “signal extraction” counter-explanation for the Dunning-Kruger effect in self-evaluation of performance. *Journal of Economic Psychology*, *39*, 85–100. https://doi.org/10.1016/j.joep.2013.07.004

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in Personnel Selection: Factor Structure in Applicant and Nonapplicant Populations. *Journal of Applied Psychology*, *78*(6), 966–974. https://doi.org/10.1037/0021-9010.78.6.966

Schmitt, M. J., & Steyer, R. (1993). A latent state-trait model (not only) for social desirability. *Personality and Individual Differences*, *14*(4), 519–529.

Schneider, D. J., & Turkat, D. (1975). Self-presentation following success or failure: Defensive self-esteem models. *Journal of Personality*.

Schoderbek, P. P., & Deshpande, S. P. (1996). Impression management, overclaiming, and perceived unethical conduct: The role of male and female managers. *Journal of Business Ethics*, *15*(4), 409–414. https://doi.org/10.1007/BF00380361

Schubert, A. L., & Frischkorn, G. T. (2020). Neurocognitive Psychometrics of Intelligence: How Measurement Advancements Unveiled the Role of Mental Speed in Intelligence Differences. *Current Directions in Psychological Science*, *29*(2), 140–146. https://doi.org/10.1177/0963721419896365

Schwaba, T., Robins, R. W., Grijalva, E., & Bleidorn, W. (2019). Does openness to experience matter in love and work? Domain, facet, and developmental evidence from a 24-year longitudinal study. *Journal of Personality*, *87*(5), 1074–1092.

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *The American Journal of Evaluation*, *22*(2), 127-160.

Schwarz, N., & Wellens, T. (1997). Cognitive dynamics of proxy responding: The diverging perspectives of actors and observers. JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-, 13, 159-180.

Seashore, R. H. (1939). Work methods: an often neglected factor underlying individual differences. *Psychological Review*, *46*(2), 123–141. https://doi.org/10.1037/h0055373

Sedikides, C. (1993). Personality Processes and Individual Differences Assessment , Enhancement , and Verification Determinants of the Self-Evaluation Process. *Journal of Personality and Social Psychology*, *65*(2), 317–338.

Sedikides, C., Gaertner, L., & Toguchi, Y. (2003). Pancultural Self-Enhancement. *Journal of Personality and Social Psychology*, *84*(1), 60–79. https://doi.org/10.1037/0022-3514.84.1.60

Sedikides, C., & Green, J. D. (2000). On the self-protective nature of inconsistency-negativity management: Using the person memory paradigm to examine self-referent memory. *Journal of Personality and Social Psychology, 79*(6), 906–922. https://doi.org/10.1037/0022-3514.79.6.906

Sedikides, C., & Gregg, A. P. (2003). Portraits of the self. *The SAGE Handbook of Social Psychology*, 93–122. https://doi.org/10.4135/9781848608221.n5

Sedikides, C., Herbst, K. C., Hardin, D. P., & Dardis, G. J. (2002). Accountability as a deterrent to self-enhancement: The search for mechanisms. *Journal of Personality and Social Psychology*, *83*(3), 592.

Sedikides, C., Hoorens, V., & Dufner, M. (2015). Self-enhancing self-presentation: Interpersonal, relational, and organizational implications. [in:] F. Guay, D. M. McInerney, R. Craven, & H. W. Marsh (Eds.), *Self-concept, motivation and identity: Underpinning success with research and practice* (pp. 29-55). Charlotte, NC: Information Age Publishing.

Sedikides, C., & Skowronski, J. J. (2000). *On the evolutionary functions of the symbolic self: The emergence of self-evaluation motives.* In A. Tesser, R. B. Felson, & J. M. Suls (Eds.), *Psychological perspectives on self and identity* (p. 91–117). American Psychological Association. https://doi.org/10.1037/10357-004

Sedikides, C., & Strube, M. J. (1997). *Self-evaluation: To thine own self be good, to thine own self be sure, to thine own self be true, and to thine own self be better.* In M. P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 29* (p. 209–269). Academic Press. https://doi.org/10.1016/S0065-2601(08)60018-0

Seed, P. (2002). CI2: Stata module to compute confidence intervals for correlations. https://ideas.repec.org/c/boc/bocode/s423603.html

Seeman, M. (1963). Alienation and social learning in a reformatory. *American Journal of Sociology*, *69*(3), 270–284.

Severo, M. C., Paul, K., Walentowska, W., Moors, A., & Pourtois, G. (2020). Neurophysiological evidence for evaluative feedback processing depending on goal relevance. *NeuroImage*, 116857.

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2013). Honesty requires time—a reply to Foerster et al.(2013). *Frontiers in psychology*, *4*, 634.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475.

Sheldon, K. M., Ryan, R. M., Rawsthorne, L. J., & Ilardi, B. (1997). Trait self and true self: Cross-role variation in the big-five personality traits and its relations with psychological authenticity and subjective well-being. *Journal of Personality and Social Psychology*, *73*(6), 1380–1393. https://doi.org/10.1037/0022-3514.73.6.1380

Shi, Y., Sedikides, C., Cai, H., Liu, Y., & Yang, Z. (2017). Disowning the self: The cultural value of modesty can attenuate self-positivity. *Quarterly Journal of Experimental Psychology*, *70*(6), 1023–1032. https://doi.org/10.1080/17470218.2015.1099711

Siedlecka, M., Skóra, Z., Paulewicz, B., Fijałkowska, S., Timmermans, B., & Wierzchoń, M. (2019a). Responses improve the accuracy of confidence judgements in memory tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 712.

Siedlecka, M., Wereszczyński, M., Paulewicz, B., & Wierzchoń, M. (2019b). Visual awareness judgments are sensitive to the outcome of performance monitoring. *BioRxiv*, *August*, 572503. https://doi.org/10.1101/572503

Silber, H., Danner, D., & Rammstedt, B. (2019). The impact of respondent attentiveness on reliability and validity. *International Journal of Social Research Methodology*, *22*(2), 153–164. https://doi.org/10.1080/13645579.2018.1507378

Simsek, Z., & Veiga, J. F. (2001). A primer on internet organizational surveys. *Organizational research methods*, *4*(3), 218-235.

Sirken, M. G. (1970). Household surveys with multiplicity. *Journal of the American statistical Association*, *65*(329), 257-266.

M.Sitek (red.) (2019). *Program Międzynarodowej Oceny Umiejętności Uczniów. Wyniki badania PISA 2018 w Polsce*. Warszawa: Instytut Badań Edukacyjnych.

Siuta, J. (1989). Zmienna aprobaty społecznej w badaniach nad zjawiskami hipnotycznymi. *Zeszyty Naukowe UJ*, *5*, 131-141.

Smith, H. (1997). *The structure and utility of social desirability scales in psychological research.* Unpublished doctoral thesis. University of Illinois, Urbana-Champaign, IL, USA. http://www.il.proquest.com/umi/

Smith, P. B. (2004). Acquiescent response bias as an aspect of cultural communication style. *Journal of Cross-Cultural Psychology*, 35, 50–61.

Smith, M. K., Trivers, R., & von Hippel, W. (2017). Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, *63*(February), 93–101. https://doi.org/10.1016/j.joep.2017.02.012

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30*(4), 526–537. https://doi.org/10.1037/h0037039

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, *31*(1), 137-149.

Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence*, *25*(2), 93–109.

Stankov, L., Lee, J., & von Davier, M. (2018). A note on construct validity of the anchoring method in PISA 2012. *Journal of Psychoeducational Assessment*, *36*(7), 709-724.

Stanovich, K. E., & Cunningham, A. E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition*, *20*(1), 51–68. https://doi.org/10.3758/BF03208254

Stanovich, K. E., & West, R. F. (1989). Exposure to Print and Orthographic Processing. *Reading Research Quarterly*, *24*(4), 402. https://doi.org/10.2307/747605

Steger, D., Schroeders, U., & Wilhelm, O. (2020). Caught in the Act : Predicting Cheating in Unproctored Knowledge. *Assessment*. https://doi.org/10.1177/1073191120914970

Steinmetz, H. C. (1932). Measuring ability to fake occupational interest. *Journal of Applied Psychology, 16*(2), 123–130. https://doi.org/10.1037/h0073177

Stephens, A. N., & Ohtsuka, K. (2014). Cognitive biases in aggressive drivers: Does illusion of control drive us off the road? *Personality and Individual Differences*, *68*, 124–129.

Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent Validity, Discriminant Validity, and Relationship with Age. *European Journal of Psychological Assessment*, *17*(3), 222–232. https://doi.org/10.1027//1015-5759.17.3.222

Stöber, J., Dette, D. E., & Musch, J. (2002). Comparing continuous and dichotomous scoring of the balanced inventory of desirable responding. *Journal of Personality Assessment*, *78*(2), 370–389. https://doi.org/10.1207/S15327752JPA7802_10

Stocké, V. (2014). Deutsche Kurzskala zur Erfassung des Bedürfnisses nach sozialer Anerkennung. Zusammenstellung sozialwissenschaftlicher Items und Skalen. *ZIS*, doi:10.6102/zis159.

Stodel, M. (2015). But what will people think?: getting beyond social desirability bias by increasing cognitive load. *International journal of market research*, *57*(2), 313-322.

Strahan, R., & Gerbasi, K. C. (1972). Short, homogeneous versions of the Marlow-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, *28*(2), 191–193. https://doi.org/10.1002/1097-4679(197204)28:2<191::AID-JCLP2270280220>3.0.CO;2-G

Stucky, B. D. & Edelen, M. O. (2015). Using hierarchical IRT models to create unidimensional measures from multidimensional data. [In:] S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment*, 183-206. New York: Routledge.

Suszek, H., Fronczyk, K., Kopera, M., Maliszewski, N., & Łyś, E. A. (2018). Psychometric properties of the Polish version of the Self-Concept Clarity Scale (SCCS). *Current Issues in Personality Psychology*, *6*(3), 181–187. https://doi.org/10.5114/cipp.2018.75842

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, *47*(2), 143–148.

Swami, V., Papanicolaou, A., & Furnham, A. (2011). Examining mental health literacy and its correlates using the overclaiming technique. *British Journal of Psychology*, *102*(3), 662–675. https://doi.org/10.1111/j.2044-8295.2011.02036.x

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361-370.

Swann, W. B., & Read, S. J. (1981a). Self-verification processes: How we sustain our self-conceptions. *Journal of Experimental Social Psychology*, *17*(4), 351–372.

Swann, W. B., & Read, S. J. (1981b). Acquiring self-knowledge: The search for feedback that fits. *Journal of Personality and Social Psychology*, *41*(6), 1119.

Swann, WB, J. (1990). To be adored or to be known? The interplay of self-enhancement and self-verification. *Foundations of Social Behavior*, *January 1990*, 408–448. http://psycnet.apa.org/psycinfo/1990-98254-012

Szpitalak, M. (2012). *Motywacyjne mechanizmy efektu dezinformacji*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Szpitalak, M., & Polczyk, R. (2015). *Samoocena: geneza, struktura, funkcje i metody pomiaru*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Sztabiński, F. 2011. *Ocena jakości danych w badaniach surveyowych.* Warszawa: Wydawnictwo IFiS PAN.

Tan, L., & Grace, R. C. (2008). Social desirability and sexual offenders: A review. *Sexual Abuse: Journal of Research and Treatment*, *20*(1), 61–87. https://doi.org/10.1177/1079063208314820

Tanaka, J. S. (1987). " How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child development, 58 (1),* 134-146.

Tao, P., Guoying, D., & Brody, S. (2009). Preliminary study of a Chinese language short form of the Marlowe–Crowne social desirability scale. *Psychological reports*, *105*(3_suppl), 1039-1046.

Taylor, S. E. (1982). The Availability Bias in Social Perception and Interaction. [In:] *Judgment under Uncertainty: Heuristics and Biases*, Daniel Kahneman. Paul Slovic, and Amos Tversky (eds.). New York: Cambridge University Press.

Taylor, S. E., & Armor, D. A. (1996). Positive illusions and coping with adversity. *Journal of Personality*, *64*(4), 873–898.

Taylor, S. E., & Brown, J. D. (1988). Illusion and Well-Being: A Social Psychological Perspective on Mental Health. *Psychological Bulletin*, *103*(2), 193–210. https://doi.org/10.1037/0033-2909.103.2.193

Tendeiro, J. N. (2018). Package 'PerFit'. https://cran.stat.unipd.it/web/packages/PerFit/PerFit.pdf

Teovanović, P., Knežević, G., & Stankov, L. (2015). Individual differences in cognitive biases: Evidence against one-factor theory of rationality. *Intelligence*, *50*, 75-86.

Tesser, A., & Rosen, S. (1975). The reluctance to transmit bad news. In *Advances in experimental social psychology* (Vol. 8, pp. 193-232). Academic Press.

Tett, R. P., & Simonet, D. V. (2011). Faking in personality assessment: A "Multisaturation" perspective on faking as performance. *Human Performance*, *24*(4), 302–321. https://doi.org/10.1080/08959285.2011.597472

Thomaes, S., Brummelman, E., & Sedikides, C. (2017). Why Most Children Think Well of Themselves. *Child Development*, *88*(6), 1873–1884. https://doi.org/10.1111/cdev.12937

Thomas, A. K., Bulevich, J. B., & Dubois, S. J. (2012). An analysis of the determinants of the feeling of knowing. *Consciousness and Cognition*, *21*(4), 1681–1694. https://doi.org/10.1016/j.concog.2012.09.005

Thompson, S. C. (1999). Illusions of control: How we overestimate our personal influence. *Current Directions in Psychological Science*, *8*(6), 187–190. https://doi.org/10.1111/1467-8721.00044

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, *4*(1), 25–29. https://doi.org/10.1037/h0071663

Thornton, G. C. (1980). Psychometric Properties of Self-Appraisals of Job Performance. *Personnel Psychology*, *33*(2), 263–271. https://doi.org/10.1111/j.1744-6570.1980.tb02348.x

Tillman, C. M., & Wiens, S. (2011). Behavioral and ERP indices of response conflict in Stroop and flanker tasks. *Psychophysiology*, *48*(10), 1405-1411.

Tobias, S., & Everson, H. T. (2002). *Knowing What You Know and What You Don't: Further Research on Metacognitive Knowledge Monitoring.* Research Report No. 2002-3. College Entrance Examination Board.

Tonković, M., Galić, Z., & Jerneić, Ž. (2011). The construct validity of over-claiming as a measure of egoistic enhancement. *Review of Psychology*, *18*(1), 13–21.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. Psychological Bulletin, 103(3), 299–314. https://doi.org/10.1037/0033-2909.103.3.299

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Tracy, J. L., Cheng, J. T., Robins, R. W., & Trzesniewski, K. H. (2009). Authentic and hubristic pride: The affective core of self-esteem and narcissism. *Self and Identity*, *8*(2–3), 196–213. https://doi.org/10.1080/15298860802505053

Trapnell, P. D., & Paulhus, D. L. (2012). Agentic and communal values: Their scope and measurement. *Journal of Personality Assessment*, *94*(1), 39–52.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, *93*(3), 239.

Trope, Y., & Neter, E. (1994). Reconciling competing motives in self-evaluation: the role of self-control in feedback seeking. *Journal of Personality and Social Psychology*, *66*(4), 646-657. https://doi.org/10.1037/0022-3514.66.4.646

Turcu, R. (2011). *The effects of warnings on applicant faking behavior.* Unpublished Master Thesis, California State University, Sacramento, CA.

Turner, J. C. (1975). Social comparison and social identity: Some prospects for intergroup behaviour. *European Journal of Social Psychology*, *5*(1), 1–34.

Tzelgov, J., & Henik, A. (1991). Suppression Situations in Psychological Research: Definitions, Implications, and Applications. *Psychological Bulletin*, *109*(3), 524–536. https://doi.org/10.1037/0033-2909.109.3.524

Urbán, R., Szigeti, R., Kökönyei, G., & Demetrovics, Z. (2014). Global self-esteem and method effects: Competing factor structures, longitudinal invariance, and response styles in adolescents. *Behavior Research Methods*, *46*(2), 488–498. https://doi.org/10.3758/s13428-013-0391-5

Uziel, L. (2010). Rethinking social desirability scales: From impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, *5*(3), 243–262. https://doi.org/10.1177/1745691610369465

Uziel, L. (2014). Impression management ("Lie") scales are associated with interpersonally oriented self-control, not other-deception. *Journal of Personality*, *82*(3), 200–212. https://doi.org/10.1111/jopy.12045

van der Linden, D., Dunkel, C. S., & Petrides, K. V. (2016). The General Factor of Personality (GFP) as social effectiveness: Review of the literature. *Personality and Individual Differences*, *101*, 98–105. https://doi.org/10.1016/j.paid.2016.05.020

van Dijk, T. K., Datema, F., Welten, S. C. M., & van de Vijver, F. J. R. (2009). Acquiescence and extremity in cross-national surveys: Domain dependence and country-level correlates. In G. Aikaterini & K. Mylonas (Eds.), *Quod Erat Demonstrandum: From Herodotus' ethnographic journeys to cross-cultural research: Proceedings from the 18th International Congress of the International Association for Cross-Cultural Psychology.* https://scholarworks.gvsu.edu/iaccp_papers/51/

van Hooft, E. A. J., & Born, M. P. (2012). Intentional response distortion on personality tests: Using eye-tracking to understand response processes when faking. *Journal of Applied Psychology*, *97*(2), 301–316. https://doi.org/10.1037/a0025711

Van Lange, P. A. M., & Sedikides, C. (1998). Being more honest but not necessarily more intelligent than others: Generality and explanations for the Muhammad Ali effect. *European Journal of Social Psychology*, *28*(4), 675–680.

van Prooijen, J. W., & Krouwel, A. P. M. (2019). Overclaiming Knowledge Predicts Anti-establishment Voting. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550619862260

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

Vanyperen, N. W. (1992). Self-Enhancement Among Major League Soccer Players: The Role of Importance and Ambiguity on Social Comparison Behavior. *Journal of Applied Social Psychology*, *22*(15), 1186–1198. https://doi.org/10.1111/j.1559-1816.1992.tb02359.x

Vasilopoulos, N. L., Cucina, J. M., & McElreath, J. M. (2006). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology*, *90*(2), 306–322. https://doi.org/10.1037/0021-9010.90.2.306

Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology, 98*(2), 281–300.

Vecchione, M., & Alessandri, G. (2013). Disentangling trait from state components in the assessment of egoistic and moralistic self-enhancement. *Personality and Individual Differences*, *54*(8), 884–889. https://doi.org/10.1016/j.paid.2012.12.027

Vecchione, M., & Alessandri, G. (2014). Egoistic and moralistic self-enhancement in the eye of the beholder: A cross-informant study. *Journal of Personality*, *82*(5), 432–439. https://doi.org/10.1111/jopy.12073

Veen, V. van, & Carter, C. S. (2006). Conflict and cognitive control in the brain. *Current Directions in Psychological Science*, *15*(5), 237–240.

Velicer, W. F., & Weiner, B. J. (1975). Effects of sophistication and faking sets on the Eysenck Personality Inventory. *Psychological Reports*, *37*(1), 71–73.

Verardi, S., Dahourou, D., Ah-Kion, J., Bhowon, U., Tseung, C. N., Amoussou-Yeye, D., ... & Barry, O. (2010). Psychometric properties of the Marlowe-Crowne social desirability scale in eight African countries and Switzerland. *Journal of Cross-Cultural Psychology*, *41*(1), 19-34.

Vernon, P. E. (1934). The attitude of the subject in personality testing. *Journal of Applied Psychology, 18*(2), 165–177. https://doi.org/10.1037/h0074033

Vispoel, W. P., Morris, C. A., & Clough, S. J. (2019). Interchangeability of Results From Computerized and Traditional Administration of the BIDR: Convenience Can Match Reality. *Journal of Personality Assessment*, *101*(3), 237–252. https://doi.org/10.1080/00223891.2017.1406361

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*(2), 197–210.

von Davier, M. (2010). Why Sum Scores May Not Tell Us All About Test Takers. *Newborn and Infant Nursing Reviews*, *10*(1), 27–36. https://doi.org/10.1053/j.nainr.2009.12.011

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, *2*(1), 9–36.

Von Hippel, W., & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, *34*(1), 1-56. doi:10.1017/S0140525X10001354

Von Hoyer, J., Pardi, G., Kammerer, Y., & Holtz, P. (2019). Metacognitive judgments in searching as learning (SAL) Tasks: Insights on (Mis-) calibration, multimedia usage, and confidence. *SALMM 2019 - Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information, Co-Located with MM 2019*, 3–10. https://doi.org/10.1145/3347451.3356730

Vonkova, H., Papajoanu, O., & Stipek, J. (2018). Enhancing the Cross-Cultural Comparability of Self-Reports Using the Overclaiming Technique: An Analysis of Accuracy and Exaggeration in 64 Cultures. *Journal of Cross-Cultural Psychology*, *49*(8), 1247–1268. https://doi.org/10.1177/0022022118787042

Wagenmakers, E. J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*(5), 641-671.

Wagenmakers, E.-J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, 114(3), 830–841. https://doi.org/10.1037/0033-295X.114.3.830

Wang, Y., Kim, E. S., Dedrick, R. F., Ferron, J. M., & Tan, T. (2018). A multilevel bifactor approach to construct validation of mixed-format scales. *Educational and psychological measurement*, 78(2), 253-271.

Ward, M. K., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, *76*, 417–430. https://doi.org/10.1016/j.chb.2017.06.032

Warner, S. L. (1965). Randomized Response : A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, *60*(309), 63–69.

Watkins, D., McInerney, D., Akande, A., & Lee, C. (2003). An investigation of ethnic differences in the motivation and strategies for learning of students in desegregated South African schools. *Journal of Cross-Cultural Psychology*, *34*(2), 189–194.

Weaver, R., & Prelec, D. (2013). Creating truth-telling incentives with the Bayesian truth serum. *Journal of Marketing Research*, *50*(3), 289–302. https://doi.org/10.1509/jmr.09.0039

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods, 15*(1), 96–110. https://doi.org/10.1037/a0018721

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*(3), 409-422.

West, R. F., & Stanovich, K. E. (1991). the Incidental Acquisition of Information From Reading. *Psychological Science*, *2*(5), 325–330. https://doi.org/10.1111/j.1467-9280.1991.tb00160.x

Wetzel, E., & Greiff, S. (2018). The world beyond rating scales: Why we should think more carefully about the response format in questionnaires. *European Journal of Psychological Assessment*, *34*(1), 1–5. https://doi.org/10.1027/1015-5759/a000469

Wetzel, E., Böhnke, J. R., Brown, A. (2016). Response Biases. *The ITC International Handbook of Testing and Assessment*, 349–363. https://doi.org/10.1093/med:psych/9780199356942.003.0024

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2015). The Stability of Extreme Response Style and Acquiescence Over 8 Years. *Assessment*, *23*(3), 279–291.

Whittlesea, B. W. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1235-1253.

Whittlesea, B. W. A., & Williams, L. D. (2000). The source of feelings of familiarity: The discrepancy-attribution hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 547–565. https://doi.org/10.1037/0278-7393.26.3.547

Why, Y. P., & Huang, R. Z. (2011). Positive illusions and its association with cardiovascular functions. *International Journal of Psychophysiology*, *81*(3), 305–311. https://doi.org/10.1016/j.ijpsycho.2011.07.016

Wicklund, R. A. (1975). Objective self-awareness. In *Advances in experimental social psychology* (Vol. 8, pp. 233–275). Elsevier.

Wiggins, J. S. (1964). Convergences among stylistic response measures from objective personality tests. *Educational and Psychological Measurement*, *24*(3), 551-562.

Wiggins, J. S. (1991). Agency and communion as conceptual coordinates for the understanding and measurement of interpersonal behavior. *Thinking Clearly about Psychology: Essays in Honor of Paul E. Meehl, Vol. 2: Personality and Psychopathology*, 89–113.

Wilhelm, K., Dewhurst-Savellis, J., & Parker, G. (2000). Teacher stress? An analysis of why teachers leave and why they stay. *Teachers and Teaching*, *6*(3), 291-304.

Williams, K. M., Paulhus, D. L., & Nathanson, C. (2002). *The Nature of Overclaiming : Personality and Cognitive Factors.* Poster Presented at the 110th Annual Meeting of the American Psychological Association, Chicago, IL, August.

Wilson, T. D., & Dunn, E. W. (2004). Self-Knowledge: Its Limits, Value, and Potential for Improvement. *Annual Review of Psychology*, *55*(1), 493–518.

Wojciszke, B. (1984). Skala pragmatyzmu—treść i charakterystyka psychometryczna. *Przegląd Psychologiczny*, *27*(3), 725–743.

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology, 67*(2), 222–232. https://doi.org/10.1037/0022-3514.67.2.222

Wojciszke, B. (2002). Człowiek wśród ludzi. *Zarys psychologii społecznej [A man among people. The outline of social psychology],* Warszawa: Wydawnictwo Naukowe Scholar.

Wojciszke, B. (2011). *Psychologia społeczna*. Warszawa: Wydawnictwo Naukowe Scholar.

Wojciszke, B., Baryla, W., Parzuchowski, M., Szymkow, A., & Abele, A. E. (2011). Self-esteem is dominated by agentic over communal information. *European Journal of Social Psychology*, *41*(5), 617–627.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 189–194. https://doi.org/10.1007/s10862-005-9004-7

Woodworth, R. S. (1917). Some criticisms of the Freudian psychology. *The Journal of Abnormal Psychology*, 12(3), 174–194. https://doi.org/10.1037/h0069811

Woszczynski, A. B., & Whitman, M. E. (2004). The problem of common method variance in IS research. [in:] *The handbook of information systems research* (pp. 66-78). Igi Global.

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*, *41*(2), 257–267.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005

Xu, J., Chen, J., Zhang, W., Li, W., & Sheng, Y. (2015). Faking in Job Applicants' Responses in Personality Tests: Evidence from An Eye-Tracking Study of Job Desirability. *Acta Psychologica Sinica*, *47*(11), 1395-1404.

Yan, T., Fricker, S., & Tsai, S. (2020). Response Burden: What Is It and What Predicts It? *Advances in Questionnaire Design, Development, Evaluation and Testing*, 193–212. https://doi.org/10.1002/9781119263685.ch8

Yang, Z., Barnard-Brak, L., & Lan, W. Y. (2019). Examining the association of over-claiming with mathematics achievement. *Learning and Individual Differences*, *70*(January), 30–38. https://doi.org/10.1016/j.lindif.2019.01.004

Yentes, R., & Wilhelm, F. (2018). Careless: Procedures for computing indices of careless responding. *R package version*, *1*(3), 2018.

Yik, M. S. M., Bond, M. H., & Paulhus, D. L. (1998). Do Chinese self-enhance or self-efface? It's a matter of domain. *Personality and Social Psychology Bulletin*, *24*(4), 399–406.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. https://doi.org/10.1006/jmla.2002.2864

Yuan, Y., Bing, M., Hou, N., Zheng, L., Hack, E., Davison, H. K., Chelsea, V., & Kluemper, D. (2015). A Laboratory Investigation of Validities of Four Faking Measures. *Paper presented at the annual meeting of The Society for Industrial and Organizational Psychology, Philadelphia, PA*.

Zaborowski, Z. (1989). *Psychospołeczne problemy samoświadomości*. Warszawa: Państwowe Wydawnictwo Naukowe.

Zajenkowski, M., Czarna, A. Z., Szymaniak, K., & Dufner, M. (2019). What do highly narcissistic people think and feel about (their) intelligence? *Journal of Personality*, *October*. https://doi.org/10.1111/jopy.12520

Zawistowska, A. (2013). Płeć matematyki. *Studia Socjologiczne*, *3*(210), 75–95.

Zawistowska, A. (2017). Gender Differences in High-Stakes Maths Testing. Findings From Poland. *Studies in Logic, Grammar and Rhetoric*, *50*(1), 205–226. https://doi.org/10.1515/slgr-2017-0025

Zawistowska, A., & Sadowski, I. (2019). Filtered Out, but Not by Skill: The Gender Gap in Pursing Mathematics at a High-Stakes Exam. *Sex Roles*, *80*(11-12), 724-734.

Zell, E., & Krizan, Z. (2014). Do People Have Insight Into Their Abilities? A Metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111–125. https://doi.org/10.1177/1745691613518075

Zerbe, W. J., & Paulhus, D. L. (1987). Socially Desirable Responding in Organizational Behavior: A Reconception. *The Academy of Management Review*, *12*(2), 250. https://doi.org/10.2307/258533

Zettler, I., Hilbig, B. E., Moshagen, M., & de Vries, R. E. (2015). Dishonest responding or true virtue? A behavioral test of impression management. *Personality and Individual Differences*, *81*, 107–111. https://doi.org/10.1016/j.paid.2014.10.007

Zhang, J., & Ziegler, M. (2015). Interaction Effects between Openness and Fluid Intelligence Predicting Scholastic Performance. *Journal of Intelligence*, *3*(3), 91–110. https://doi.org/10.3390/jintelligence3030091

Zhang, J., Paulhus, D. L., & Ziegler, M. (2018). Personality predictors of scholastic cheating in a Chinese sample. *Educational Psychology*, *0*(0), 1–19. https://doi.org/10.1080/01443410.2018.1502414

Zheng, L. (2015). *Testing the Effectiveness and Psychology of Different Types of Pre-warnings in Reducing Applicant Faking on Personality Tests within Selection Contexts.* Unpublished Msc thesis, Auburn University, AL, USA.

Zickar, M. J., & Gibby, R. E. (2006). A history of faking and socially desirable responding on personality tests. [in:] Richard L. Griffith & Mitchell M. Peterson, *A closer examination of applicant faking behavior*, Greenwich, Connecticut, IAP, 21-42.

Zickar, M.J., & Sliter, K.A. (2012). Searching for unicorns: Item Response Theory-based solutions to the faking problem. [in:] Matthias Ziegler, Carolyn MacCann, Richard D. Roberts, *New perspectives on faking in personality assessment,* Oxford University Press.

Ziegler, M. (2011). Applicant Faking: A Look Into the Black Box. *TIP: The Industrial-Organizational Psychologist*, *49*(1), 29–36. http://www.siop.org/tip/july11/06ziegler.aspx

Ziegler, M. (2015). "F**** You, I won't do what you told me!" - Response biases as threats to psychological assessment. *European Journal of Psychological Assessment*, *31*(3), 153–158. https://doi.org/10.1027/1015-5759/a000292

Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement*, *69*(4), 548–565. https://doi.org/10.1177/0013164408324469

Ziegler, M., & Kemper, C. J. (2013). Extreme Response Style and Faking: Two Sides of the Same Coin? *Interviewers' Deviations in Surveys – Impact, Reasons, Detection and Prevention*, *January 2013*, 217–233.

Ziegler, M., Danay, E., Heene, M., Asendorpf, J., & Bühner, M. (2012). Openness, fluid intelligence, and crystallized intelligence: Toward an integrative model. *Journal of Research in Personality*, *46*(2), 173–183. https://doi.org/10.1016/j.jrp.2012.01.002

Ziegler, M., Kemper, C., & Rammstedt, B. (2013). The vocabulary and overclaiming test (VOC-T). *Journal of Individual Differences*, *34*(1), 32–40. https://doi.org/10.1027/1614-0001/a000093

Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What Is the Nature of Faking? Modeling Distinct Response Patterns and Quantitative Differences in Faking at the Same Time. *Organizational Research Methods*, *18*(4), 679–703. https://doi.org/10.1177/1094428115574518

Ziegler, M., MacCann, C., & Roberts, R. D. (2012). *New Perspectives on Faking in Personality Assessment*, 1–384. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.001.0001

Zimmerman, J., Broder, P. K., Shaughnessy, J. J., & Underwood, B. J. (1977). A recognition test of vocabulary using signal-detection measures, and some correlates of word and nonword recognition. *Intelligence*, *1*(1), 5–31. https://doi.org/10.1016/0160-2896(77)90025-3

Zinczuk, J., & Draheim, S. E. (2009). Represywny styl radzenia sobie z bodźcami zagrażającymi a samooszukiwanie. *Czasopismo Psychologiczne,* 15, 1, 23-41.

Zinczuk-Zielazna, J., & Słysz, A. (2019). Autopercepcja cech osobowości u kobiet niskolękowych, wysokolękowych i wypierających. *Annales Universitatis Mariae Curie-Skłodowska, Sectio J – Paedagogia-Psychologia*, *31*(4), 197. https://doi.org/10.17951/j.2018.31.4.197-217

Zuber, I. (1981). Characteristics of self-esteem and perceptual sensitivity to pictures of self and others. *Polish Psychological Bulletin*, *12*(2), 97–105.

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. *Advances in Social Science Methodology, 5(1),* 269–304.

Żylicz, P., & Malinowska, D. (2012). Przygotowanie polskich eksperymentalnych wersji skal samoopisowych dla nauczycieli wraz z badaniem pilotażowym. *Unpublished report commissioned by Instytut Badań Edukacyjnych.*

10- APPENDICES

These sections contain supplementary materials that expand information on results obtained in the analyses commented above. Each Appendix can be reached by clicking the below Dropbox links. Upon clicking the link a zipped folder will be downloaded on computer. The zipped folders will typically contain tables in the .xls format and figures and pictures in the .png format.

10.1 Appendix A

This Appendix contains item characteristic curves (ICCs) from the IRT scaling of the math familiarity scale described in the section 6.2. The ICCs are in .png format.

https://www.dropbox.com/sh/7lhnqfyt5qg6ch2/AAB1MEzJrZJtxVClv0QxbaW_a?dl=0

10.2 Appendix B

This package comprises .xlsx files in which basic psychometric qualities of the self-reports scales used in the work are displayed. Internal consistency and EFA results are presented. Also list of items used to create the scale is given.

https://www.dropbox.com/sh/jzw9fzxa0e7n0gz/AADphpumQEYkbWt54g4FzZ3Ma?dl=0

10.3 Appendix C

In this Appendix all additional tables that were not presented in the main body of the text are supplied. Files in the .xlsx format.

https://www.dropbox.com/sh/uqull1zj9pgn3hc/AADl5iBdyKPoq-7mTgfErUX6a?dl=0

10.4 Appendix D

This section offers ICCs from the DIF analysis described in the section 7.9.2. Pictures in .png format.

https://www.dropbox.com/sh/tpfr68p7wqabr6m/AACtAqfsIgTJFz55THGObNUHa?dl=0

10.5 Appendix E

This Appendix contains additional tables and statistics for the CFA models analysed in the section 7.7 of the work. Additionally, factor loadings diagrams are enclosed, both in .pdf and .dgm (MPlus diagrammer) format.

https://www.dropbox.com/sh/0ntnsrllx3xpzig/AABGCm7Q-WifxqcMrNywVHTla?dl=0

10.6 Appendix F

This Appendix comprises materials for the four-class solution from the latent class analysis described in subchapter 7.10: MPlus output file and MPlus plot in .jpg format.

https://www.dropbox.com/sh/pyqpirqmkxdb0zx/AAAScV2gYYNxmDU_wrRvzwMPa?dl=0